DOCUMENT RESUME

ED 454 853 IR 058 144

TITLE	Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000).
INSTITUTION	Library of Congress, Washington, DC. Cataloging Directorate.
PUB DATE	2000-11-00
NOTE	538p.; For individual papers, see IR 058 145-166.
AVAILABLE FROM	For full text: http://lcweb.loc.gov/catdir/bibcontrol/conference.html.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF02/PC22 Plus Postage.
DESCRIPTORS	*Access to Information; Authority Control (Information); *Cataloging; Electronic Libraries; *Information Services; Library Catalogs; Library Surveys; *World Wide Web
IDENTIFIERS	Anglo American Cataloging Rules 2; *Electronic Resources; Library of Congress; MARC; *Metadata; Web Sites

ABSTRACT

The goals of this conference, sponsored by the Library of Congress Cataloging Directorate, were to develop an overall strategy to address the challenges of improved access to World Wide Web resources through library catalogs and applications of metadata and to identify attainable actions for achieving the objectives of the overall strategy. This proceedings contains the text of the keynote address, "From Card Catalogues to WebPACs: Celebrating Cataloguing in the 20th Century" (Michael Gorman) and the following conference papers: "The New Context for Bibliographic Control in the New Millennium" (Clifford Lynch); "Metadata for Web Resources: How Metadata Works on the Web" (Martin Dillon); "The Catalog as Portal to the Internet" (Sarah E. Thomas); "The Library Catalogue in a Networked Environment" (Tom Delsey); "International Metadata Initiatives: Lessons in Bibliographic Control" (Priscilla Caplan); "Is Precoordination Unnecessary in LCSH? Are Web Sites More Important To Catalog than Books? A Reference Librarian's Thoughts on the Future of Bibliographic Control" (Thomas Mann); "Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources" (Matthew Beacom); "Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources: Issues and Challenges" (Lois Mai Chan); "Resource Discovery Using Z39.50: Promise and Reality" (William E. Moen); "Authority Control on the Web" (Barbara B. Tillett); "AACR2 and Its Place in the Digital World: Near-Term Solutions and Long-Term Direction" (Ann Huthwaite); "Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives" (Sally McCallum); "Business Unusual: How Event-Awareness' May Breathe Life into the Catalog?" (Carle Lagoze); "Descriptive Resource Needs from the Reference Perspective: Report on a Survey of U.S. Reference Librarians" (Carolyn Larson and Linda Arret); "Some Observations on Metadata and Digital Libraries" (Caroline R. Arms); "An Initial Survey and Description of How Selected United States Government Libraries, Information Centers, and Information Services Provide Public Access to Information via the Internet" (Thomas A. Downing); "A Comparison of Web Resource Access Experiments: Planning for the New Millennium" (Jane Greenberg); "Redesign of Library Workflows: Experimental Models for Electronic Resource Description" (Karen Calhoun); "Metadata, Cataloging, Digitization and Retrieval: Who's Doing What to Whom: the Colorado Digitization Project Experience" (Liz Bishoff and

Reproductions supplied by EDRS are the best that can be made
from the original document.

William A. Garrison); "Exploring Partnerships: What Can Producers and Vendors Provide?" (Michael Kaplan); and "Partnerships to Mine Unexploited Sources of Metadata" (Regina Romano Reynolds). The conference program is also included. (MES)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 454 853

Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000)

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

B. Wiggins

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR058144

BEST COPY AVAILABLE

What's new

Topical Discussion group recommendations

Cybercast information

Candid photos from the conference

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Topical Discussion group recommendations are available
Please send comments to jbyr@loc.gov

Cybercasts of speakers and presentations are available.

DATES:

November 15-17, 2000

PLACE:

Library of Congress, accommodating approximately 125 attendees (including invited speakers and panelists, and participants), with consideration given to cybercasting. General sessions to be held in the Mumford Room; topical discussion groups in conference rooms.

INTENDED AUDIENCE:

Librarians, who are versed in the use of AACR2 and metadata information schemes, including those providing reference and computer-based information services; metadata developers involved in applying metadata to Web resources; computer and other information specialists actively engaged in creating software tools to access Web content; library vendors developing next-generation WebPACS; Web authors and producers designing content for improved access.

PURPOSE AND GOALS:

To celebrate the Library of Congress's Bicentennial and its historic and outstanding role in providing national and international leadership to the library profession in the development of cataloging policy and to the library

Logistical Information for conference participants

Conference Organizing Team

Cataloging Directorate Home Page

Library of Congress Home Page

community in its production of standardized records to enable bibliographic control and access to resources in a variety of formats. It is the aim of this conference to bring together authorities in the cataloging and metadata communities to discuss outstanding issues involving improved discovery and access to Web resources within the framework of international standards. The focus of the conference is on an open discussion of the issues with primary attention on proposed solutions and action items which result. For this reason, presentations and panels are not intended merely to convey or update existing information but to frame the issues and fashion solutions to problems. The conference will produce recommendations that will help the Library of Congress, the framers of AACR, and the library profession develop and implement an effective response to the bibliographic challenges posed by the proliferation of Web resources.

The goals of the conference are:

1. To develop an overall strategy to address the challenges of improved access to Web resources through library catalogs and applications of metadata. Specifically: a) Plan a national agenda that includes identifying library resource description needs and future directions of the library catalog in the Web environment; b) Promote changes to AACR2 that are coherent, flexible, and adaptable to accessing the proliferation and diversity of Web resources; c) Encourage wider use of authorised subject and classification systems, such as LCSH, LCC, and DDC, to enhance resource organization and discovery; d) Collaborate with metadata communities to develop and refine metadata element sets that support interoperability between different systems based on different metadata; e) Participate in developing and promoting national and international standards that will enable libraries and metadata communities to meet the new and changing needs of Web users; f) Foster the development of software (templates, intelligent agents) for use in generating library resource descriptions embedded in or linked to the resources described; g) Identify and address related training issues and needs; and, h) Support the development of mechanisms to facilitate efficient interfaces between the library catalog and other sources of metadata on the Web.
2. To identify attainable actions for achieving the objectives of the overall strategy. Such actions would include those resulting from goals 1a)-g) noted above. These actions could have several outcomes, ranging from the initiation of new or expanded bibliographic projects to the development of partnerships among representatives of the library, metadata, and vendor communities in attendance. It is expected that Conference participants will frame a plan for moving recommendations and actions forward to their respective organizations and other groups sharing common concerns.

DESCRIPTION:

In the course of the last five years, libraries have witnessed an explosion in the quantity of digital resources that have become available on the World Wide Web. These materials comprise a bibliographical mix of known types or genres (serials and other text-based items) and newer forms such as multimedia, home pages, databases, discussion forums, and online services. Within this period, libraries began to recognize the importance of digital resources and the need to make them an integral part of their collections. However, these resources have presented a number of cataloging problems related to their bibliographic control. Such problems involve content, format, and technology issues which have resulted in raising questions about the overall ability of established cataloging practice as embodied in the Anglo-American Cataloguing Rules (AACR2), and in the application of traditional library subject and classification tools, such as the Library of Congress Subject Headings (LCSH), Library of Congress Classification (LCC) and Dewey Decimal Classification (DCC), to deal with these materials. As a consequence, various groups within the national cataloging community have undertaken separate but related, and in some cases, overlapping initiatives to address these problems.

At the same time, new metadata information schemes have been developed promising greater precision in the discovery and access to Web resources. Prominent among these schemes are the Dublin Core (DC), the Encoded Archival Description (EAD), and the Text Encoding Initiative (TEI). In tandem with the development of metadata schemes, there are a number of national and international projects underway that are exploring the creation and use of metadata, primarily for Web resources. Among these are OCLC's CORC (Cooperative Online Resource Catalog), the Nordic Metadata Project, and BIBLINK, to name just a few.

These different and diverse developments underscore the need to bring together leaders in the library and other metadata communities to discuss their work and to share their goals and contributions. This special Conference provides that opportunity with a program dedicated to the theme of promoting the effective organization of networked resources.

CONFERENCE FORMAT:

Speakers will present formal summaries of key points and recommendations in contributed papers, which will be made available on an open electronic discussion list in advance of the Conference. All participants will be expected to read these papers in advance and offer feedback through contributions to the discussion list; the speakers will consider this feedback in preparing their presentations. Panelists will also contribute papers that they will summarize at the Conference. Speakers, panelists, and those summarizing session highlights will help to identify key topics and issues for participants to address in their recommendations and action plans. Topical discussion groups will further the effort to seek solutions to problems and derive action items, and to facilitate discussion on issues. It is anticipated that this format will engender energetic discussion of the theoretical and practical issues. Assembled conferees will consider the recommendations of the topical discussion groups and help to develop and prioritize them into a strategic plan.

CONFERENCE PROGRAM:

Sessions and panel discussions address selected topics bearing on the conference theme. As currently projected, these include:

Keynote Address

From Card Catalogues to WebPACS: Celebrating Cataloguing in the 20th Century.
Speaker: Michael Gorman

Dinner Speaker

The New Context for Bibliographic Control In the New Millennium
Speaker: Clifford Lynch

The Library Catalog and the Web

Discussion Paper:

Metadata for Web Resources: How Metadata Works on the Web
Author: Martin Dillon

Conference Presentations:

The Catalog as Portal to the Internet
Speaker: Sarah Thomas
Commentator: Brian Schottlaender

The Library Catalogue in a Networked Environment
Speaker: Tom Delsey
Commentator: Jennifer Trant

International Metadata Initiatives: Lessons in Bibliographic Control
Speaker: Priscilla Caplan
Commentator: Robin Wendler

Assessing Current Library Standards for Bibliographic Control and Web Access

Discussion paper:

Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books? : A Reference Librarian's Thoughts on the Future of Bibliographic Control
Author: Thomas Mann

Crossing a Digital Divide: AACR2 and Unaddressed Problems of

Networked Resources
Speaker: Matthew Beacom
Commentator: Glenn Patton

Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources
Speaker: Lois Mai Chan
Commentator: Diane Vizine-Goetz

Resource Discovery Using Z39.50: Promise and Reality
Speaker: William E. Moen

Authority Control on the Web
Speaker: Barbara Tillett

Future Directions

AACR2 and Its Place in the Digital World: Near-term Revisions and Long-term Direction
Speaker: Ann Huthwaite
Commentator: Lynne Howarth

Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives
Speaker: Sally McCallum
Commentator: Paul Weiss

Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?
Speaker: Carl Lagoze

Descriptive Resource Needs from the Reference Perspective
Speakers: Linda Arret, Carolyn Larson

Experimentation

Discussion Papers:

Some Observations on Metadata and Digital Libraries
Author: Caroline Arms

An Initial Survey and Description of How Selected United States Government Libraries, Information Centers, and Information Services Provide Public Access to Information Via the Internet
Author: Thomas Downing

Conference Presentations:

A Comparison of Web Resource Access Experiments: Planning for the New Millennium
Speaker: Jane Greenberg

Redesign of Library Workflows: Experimental Models for Electronic Resource Description
Speaker: Karen Calhoun

Exploring Partnerships

Discussion Paper:

Metadata, Cataloging, Digitization and Retrieval: Who's Doing What to Whom: The Colorado Digitization Project Experience
Authors: Liz Bishoff, Bill Garrison

Conference Presentations:

Exploring Partnerships: What Can Producers and Vendors Provide?
Speaker: Michael Kaplan
Commentators: Amira Aaron, Jeff Calcagno, Lynn Connaway

Partnerships to Mine Unexploited Sources of Metadata
Speaker: Regina Reynolds

Outcomes

Development, Completion and Presentation of Action Plans

November 3, 2000

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

What's New on the Conference Web Site

All final versions of papers and commentaries are up as of January 31, 2001

Topical Discussion group recommendations January 04, 2001

All Cybercasts of the speakers are now available.

Logistical information for participants last updated 10/31/00

Papers:

11/8/00 Digests of comments submitted to the BibControl listserv are available. One digest deals with general comments about the conference and ideas presented in the papers, the other deals with comments about specific papers

11/7/00 Metadata, Cataloging, Digitization and Retrieval: Who's Doing What to Whom: The Colorado Digitization Project Experience
Authors: Liz Bishoff and Bill Garrison

11/2/00 An Initial Survey and Description of How Selected United States Government Libraries, Information Centers, and Information Services Provide Public Access to Information Via the Internet
Author: Thomas A. Downing

11/2/00 From Card Catalogues to WebPACS: Celebrating Cataloguing in the 20th Century
Author: Michael Gorman

Cataloging Directorate Home Page

Library of Congress Home Page

10/31/00 Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books? : A Reference Librarian's Thoughts on the Future of Bibliographic Control
Author: Thomas Mann

10/31/00 The sidebar link to the NAS report has been changed to take you to a page that includes 2 articles from the LC staff paper, the *Gazette*.

10/26/00 Descriptive Resource Needs from the Reference Perspective
Authors: Carolyn Larson and Linda Arret

10/23/00 All conference participants may want to familiarize themselves with the National Academy of Sciences report, LC21: A Digital Strategy for the Library of Congress (http://www.nap.edu/books/0309071445/html/)

10/18/00 Partnerships to Mine Unexploited Sources of Metadata
Author: Regina Reynolds

10/18/00 Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?
Author: Carl Lagoze

10/16/00 Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources
Author: Matthew Beacom

10/02/00 A Comparison of Web Resource Access Experiments: Planning for the New Millennium
Author: Jane Greenberg

10/02/00 Some Observations on Metadata and Digital Libraries
Author: Caroline R. Arms

9/28/00 Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives
Author: Sally McCallum

9/11/00 The Catalog as Portal to the Internet
Author: Sarah Thomas

8/31/00 Metadata for Web Resources: How Metadata Works on the Web
Author: Martin Dillon

8/17/00 Summary for dinner speaker Clifford Lynch's paper The New Context for Bibliographic Control In the New Millennium

8/16/00 Redesign of Library Workflows: Experimental Models for Electronic Resource Description
Author: Karen Calhoun

8/16/00 Summary for discussion paper Metadata, Cataloging, Digitization and Retrieval: Who is doing what to whom? The Colorado Digitization Project Experience
Authors: Liz Bishoff, Bill Garrison

8/10/00 Authority Control on the Web
Author: Barbara Tillett

8/3/00 Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources
Author: Lois Mai Chan

8/3/00 Resource Discovery Using Z39.50: Promise and Reality
Author: William E. Moen

7/24/00 Exploring Partnerships: What Can Producers and Vendors Provide?
Author: Michael Kaplan

7/18/00 The Library Catalogue in a Networked Environment
Author:Tom Delsey

7/5/00 International Metadata Initiatives: Lessons in Bibliographic Control
Author: Priscilla Caplan

6/30/00 AACR2 and Its Place in the Digital World: Near-term Revisions and Long-term Direction
Author: Ann Huthwaite

10/31/00 Participants directory is now available in PDF format

10/26/00 Communication page established providing information on internet access for participants while attending the conference, message center information, and the conference emergency number

10/26/00 Program page adjusted to move LC dinner to day 2 and adjust schedule accordingly

10/24/00 Cybercast page initiated. This is where the cybercasts will be made available.

10/23/00 In preparation for the delayed cybercast presentations, interested parties may want to view a test cybercast of an LC staff presentation on the use of CORC, the BEOnline+ and CECites+ projects as well as CORC pathfinders at LC. It is possible that your browser or some of its plug-in components may need upgrading to view the cybercast and your browser should inform you if this is the case.

10/23/00 Topical discussion groups added to the sidebar for all pages.

10/23/00 Conference attendees may choose to stay within the Madison Building during the no-host lunch or may want to get a snack during a conference break. A list of food service locations in the Madison Building is provided for your convenience. A more thorough list of places to eat city-wide will be provided in the registration packets for your free evenings.

9/1/00 Logistics page updated to add directions for driving to the conference hotel and directions from the Metro to the hotel.

8/16/00 Logistics page updated to add information about conference entrance procedures, registration, and to add maps

7/24/00 Conference program is now available.

7/13/00 Information about hotel reservations for conference participants is available on the logistics page.

7/12/00 The discussion list is up and running. It will foster discussion regarding the conference papers and the issues they address. It will provide interested parties with a means for sharing constructive feedback on the topics. This discussion list is intended to encourage interested colleagues throughout the world -- particularly those who could not be invited to attend in person due to logistical constraints and other considerations -- to participate by commenting on conference issues. Presenters and Commentators have been asked to monitor this discussion and to take into account important points you raise when developing the final versions of their papers. To subscribe, send a message to listserv@loc.gov with the message "subscribe bibcontrol [your name]".

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Greetings from the Director for Cataloging, Beacher J.E. Wiggins

"Welcome to our Web site! We in the Cataloging Directorate of the Library of Congress are keenly excited about our upcoming Bicentennial Conference. As you will see from the information contained here, we view this event as a timely and important one that will focus on bringing bibliographic control to a burgeoning body of information in a volatile format. You are invited for return visits to our site in the intervening months until the conference is convened in November."

"We will update information on the site to keep members of interested communities apprised of our plans. As the papers for the six sessions are mounted, we are especially interested in receiving your comments, which we will use as part of our deliberations at the conference as the attendees strategize on next steps. We will be establishing an electronic discussion list to provide you with an opportunity for input and to comment on the views of others. Please spread the word about our Web site and revisit it often in the coming months and help us tackle this challenging area of bibliographic control."

BEST COPY AVAILABLE

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Topical Discussion Groups

Guidelines for Topical Discussion Group Facilitators

Guidelines for Topical Discussion Group Recorders

Topical discussion group:	Meeting room:
1. Choosing Electronic Resources: What Is a Valuable Web Resource? Discussion facilitator: Olivia Madison	LM620
2. What Are The Continuing Education Needs of Professional Catalogers? Discussion facilitator: Sheila Intner	LM632
3a. What Near-Term Cooperative Partnerships Should Libraries Explore in the Digital World? Discussion facilitator: Larry Alford	LM642
3b. What Long-Term Cooperative Partnerships Should Libraries Explore in the Digital World? Discussion facilitator: William Gosling	LM501
4a. How Can *AACR2* Become More Responsive to Cataloging Networked Resources on the Web in the Near-Term? Discussion facilitator: Sherry Kelley	LM507
4b. How Can AACR2 Become More Responsive to Cataloging Networked Resources on the Web in the Long-Term? Discussion facilitator: Carlen Ruschoff	LM513

Cataloging Directorate Home Page

Library of Congress Home Page

5. What Can the Library Community Offer in Support of Semantic Interoperability? Discussion facilitator: Mary Charles Lasater	LM515
6. What Automated Tools Could Assist Libraries to Meet the Information Needs of Their Users? Discussion facilitator: Robert Wolven	LM527
7. What Steps Can The Library Take to Achieve Integrated Access to the Catalog and Other Discovery Tools? Discussion facilitator: Sherry Vellucci	LM541
8. How Can Libraries Participate More Actively in the Development of Metadata Standards? Discussion facilitator: Sally Sinn	LM542
9. How Can Catalogers and Metadata Providers Ensure that Resource Descriptions Meet Reference Needs? Discussion facilitator: Amy Tracy Wells	LM453 (I&R)

Library of Congress
November 1, 2000
Comments: lcweb@loc.gov

- What's new
- Greetings from the Director for Cataloging
- Topical discussion groups
- *NAS study and 2 articles from the LC staff *Gazette*
- Conference program
- Speakers, commentators, and papers
- Conference sponsors
- Conference discussion list
- Logistical information for conference participants
- Conference Organizing Team
- Cataloging Directorate Home

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

LC21: A Digital Strategy for the Library of Congress

Articles from the LC staff *Gazette:*

NAS: Library Needs a Digital Strategy
by Gail Fineberg

Panel Chair Briefs Staff on NAS Report
by Gail Fineberg

BEST COPY AVAILABLE

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Cataloging Directorate Home Page

Library of Congress Home Page

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Program

revised 10/31/00 to add Thomas Mann discussion paper under Topic 2
revised 10/25/00 to change date of Gala Dinner, move Topic 3B to the first day, and adjust all times for Day 2
revised 9/11/00

Day 1 Wednesday November 15, 2000		
8:00-8:30	Arrival and registration	All participants
8:30-9:00	Coffee/Continental breakfast	
9:00-9:15	Welcome, introductions, etc.	Winston Tabb, Beacher Wiggins, John Byrum
9:15-9:45	From Card Catalogues to WebPACS: Celebrating Cataloging in the 20th Century	Michael Gorman
Topic 1: The Library Catalog and the Web Discussion paper: Metadata for Web Resources: How Metadata Works on the Web Martin Dillon		
9:45-10:05	Topic 1A: The Catalog as Portal to the Internet	Sarah Thomas
10:05-10:25	Topic 1B: The Library Catalogue in a Networked Environment	Tom Delsey
10:25-10:55	Break/Coffee service	
10:55-11:15	Topic 1C: International Metadata Initiatives: Lessons in Bibliographic Control	Priscilla Caplan
11:15-11:45	Panel reactions and questions (10 minutes per panelist to review highlights and ask questions of presenter)	
	Topic 1A: Panel reactor	Brian Schottlaender
	Topic 1B: Panel reactor	Jennifer Trant
	Topic 1C: Panel reactor	Robin Wendler
11:45-12:00	Q&A	All participants
12:00-1:00	Lunch	

Topic 2: Assessing Current Library Standards for Bibliographic Control and Web Access Discussion paper: "Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books? : A Reference Librarian's Thoughts on the Future of Bibliographic Control" by Thomas Mann		
1:00-1:20	Topic 2A: Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources	Matthew Beacom
1:20-1:40	Topic 2B: Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources	Lois Mai Chan
1:40-2:00	Panel Reactions and questions (10 min. for each panelist to review highlights and ask questions of presenter)	
	2A. Panel reactor	Glenn Patton
	2B. Panel reactor	Diane Vizine-Goetz
2:00-2:15	Q&A	All participants
2:15-2:35	Topic 2C. Resource Discovery Using Z39.50: Promise and Reality	William E. Moen
2:35-2:50	Q&A	All participants
2:50-3:20	Break/Refreshments	
3:20-3:40	Topic 2D. Authority Control on the Web	Barbara Tillett
3:40-3:55	Q&A	All participants
Topic 3: Future Directions		
3:55-4:15	Topic 3A: AACR2 and Its Place in the Digital World: Near-term Revisions and Long-term Direction	Ann Huthwaite
4:15-4:35	Topic 3B: Extending MARC to Meet New Challenges in Bibliographic Control of the Web	Sally McCallum
4:35-4:55	Panel Reactions and questions (10 min. for each panelist to review highlights and ask questions of presenter)	
	Topic 3A: Panel reactor	Lynne Howarth
	Topic 3B: Panel reactor	Paul Weiss
4:55-5:10	Q&A	All participants

Day 2 Thursday November 16, 2000		
8:00-8:45	Arrival and registration Coffee/Continental breakfast	All participants
8:45-9:05	Topic 3C: Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?	Carl Lagoze

9:05-9:20	Q&A	All participants
9:20-9:40	Topic 3D: Descriptive Resource Needs from the Reference Perspective	Carolyn Larson and Linda Arret
9:40-9:55	Q&A	All participants
9:55-10:25	Break/Coffee service	
	Topic 4: Experimentation Discussion Paper: Some Observations on Metadata and Digital Libraries Caroline Arms Discussion Paper: An Initial Survey and Description of How Selected United States Government Libraries, Information Centers, and Information Services Provide Public Access to Information Via the Internet Thomas Downing	
10:25-10:45	Topic 4A: A Comparison of Web Resource Access Experiments: Planning for the New Millennium	Jane Greenberg
10:45-11:00	Q&A	All participants
11:00-11:20	Topic 4B: Redesign of Library Workflows: Experimental Models for Electronic Resource Description	Karen Calhoun
11:20-11:35	Q&A	All participants
	Topic 5: Exploring Partnerships Discussion paper: Metadata, Cataloging, Digitization and Retrieval: Who's Doing What to Whom: The Colorado Digitization Project Experience Liz Bishoff and Bill Garrison	
11:35-11:55	Topic 5A: Exploring Partnerships: Librarians, Producers and Vendors: What Do Librarians Need?	Michael Kaplan
11:55-12:10	Q&A	All participants
12:10-1:10	Lunch	
1:10-1:25	Video	All participants
1:25-1:55	Topic 5B: Panel: What Can Producers And Vendors Provide? (10 min for each Panelist)	Lynn Connaway, Jeff Calcagno, Amira Aaron
1:55-2:10	Q&A	All participants
2:10-2:30	Topic 5C: Partnerships to Mine Unexploited Sources of Metadata	Regina Reynolds
2:30-2:45	Q&A	All participants
2:45-3:05	Break/Refreshments	
3:05-5:00	Topical discussion groups	All participants
6:30-7:30	Reception and optional visits to LC exhibits in the Great Hall vicinity	

7:30-8:45	LC hosted dinner in Great Hall Featured speaker: Clifford Lynch

Day 3 Friday November 17, 2000		
8:00-8:30	Arrival/Registration	All participants
8:30-9:00	Coffee/Continental breakfast	
9:00-10:15	Topical discussion groups formulate prioritized list of short-term/long-term recommendations for specified topics	
10:15-10:30	Break/coffee service LC recorders input group recommendations into computer	
10:30-11:45	Facilitators of the 11 topical discussion groups present prioritized list of recommendations at plenary sessions; conferees pose questions/reactions (15 min. each) LC to use lists and the discussion in determining action plan/next steps following the conference	
11:45-12:15	Hosted brief lunch break--box lunch provided	
12:15-2:00	Facilitators' presentations (continued)	
2:00	Adjournment	

Library of Congress
December 14, 2000
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Cataloging Directorate Home Page

Library of Congress Home Page

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Sponsors

Leadership Sponsors

netLibrary
www.netlibrary.com

EBSCO INFORMATION SERVICES
www-us.ebsco.com

GALE GROUP
www.gale.com

Program Sponsors

Bowker
click open the world
www.bowker.com

Participating Sponsors

epixtech
www.epixtech.com

MARCIVE, Inc.
www.marcive.com

rowecom faxon
Academic & Medical Services
www.faxon.com

Ex Libris
www.exlibris-usa.com

3M Library Systems
www.mmm.com/library

INNOVATIVE
www.iii.com

BRO DART
www.brodart.com

INGRAM
INGRAM LIBRARY SERVICES INC.
www.ingramlibrary.com

BLUE ANGEL TECHNOLOGIES
www.blueangeltech.com

VTLS Inc.
VTLS
Quality Driven . . . Caring Employees
www.vtls.com

BLACKWELL'S
BOOK SERVICES
www.blackwell.com/services/techserv/techserv.htm

WILEY
Publishers Since 1807
www.wiley.com

TLC
Solutions that Deliver
THE LIBRARY CORPORATION
www.tlcdelivers.com

HW Wilson
www.hwwilson.com

OCLC Online Computer Library Center, Inc.
Providing services to libraries around the world
www.oclc.org

The LIBRARY of CONGRESS
Cataloging Distribution Service
www.loc.gov/cds

Library of Congress
September 7, 2000
Comments: lcweb@loc.gov

BEST COPY AVAILABLE

The Library *of* Congress

SEARCH THE CATALOG | SEARCH OUR WEB SITE | ABOUT OUR SITE
GIVING OPPORTUNITIES | JOBS | TODAY IN HISTORY

COLLECTIONS & SERVICES
For Researchers, Libraries & the Public

THOMAS
Legislative Information

AMERICAN MEMORY
American History in Words, Sound & Pictures

AMERICA'S LIBRARY
Fun Site for Kids & Families

COPYRIGHT OFFICE
Forms & Information

EXHIBITIONS
An Online Gallery

THE LIBRARY TODAY
News, Events & More

HELP & FAQs
General Information

Asian Pacific American Heritage Month Celebration

Above: the interior dome of the Main Reading Room at the Library of Congress
For an online tour of the Jefferson Building, click on the dome.

101 INDEPENDENCE AVE. SE
WASHINGTON, DC 20540
(202) 707-5000

COMMENTS: lcweb@loc.gov
Please Read Our Legal Notices

COLLECTIONS & SERVICES | AMERICAN MEMORY | COPYRIGHT OFFICE | THE LIBRARY TODAY
THOMAS | AMERICA'S LIBRARY | EXHIBITIONS | HELP & FAQs

BEST COPY AVAILABLE

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Logistical Information for Conference Participants

Hotel Reservation Procedures

Updated 10/26/00 to adjust date for LC hosted dinner

Individual Reservation by Telephone

subway map, directions for driving, or directions from the metro to the hotel

Lowes L'Enfant Plaza Hotel
480 L'Enfant Plaza, SW
Washington, DC 20024
800.635.5065

A block of guest rooms have been reserved for your convenience from Tuesday, November 14th to Saturday, November 18th. The hotel is pleased to offer the government rate of $118 plus 14.5 percent tax per room, per night. All guests should make reservations directly with the hotel at 800.635.5065 or 202.484.1000 between 8:00 AM and 11:30 PM on an individual basis, identifying themselves as members of the Library of Congress Bicentennial Conference on Bibliographic Control group. All reservations need to be guaranteed by check or credit card and must be received no later than October 1, 2000--after which all resevations are subject to availability. Check-in time is 3:00 p.m.---Check-out time is 1:00 p.m.

The hotel is located at the L'Enfant Plaza metro station, on the same blue/orange subway lines as the Library of Congress (at the Capitol South

Cataloging Directorate Home Page

Library of Congress Home Page

station), just 2 stops away. Metro fare is $1.10 each way.

Direct all hotel questions to Cornelia Goode 202.707.7614 or cgoo@loc.gov

---

Conference Entrance

Conference attendees will enter the James Madison Memorial Building Independence Ave. entrance (map) at 8:00 a.m. beginning November the 15th through November the 17th. Upon entering the building, please identify yourself to the police officers and proceed to the 6th floor. Directional signs will be available to assist you.

Registration

8:00-8:30 a.m.
Exit onto the 6th floor and proceed to the registration desk to receive your name badge and Conference package. Badges must be worn at all times.

Continental Breakfast

Continental breakfast available from 8:00 - 9:00 a.m. located on the west side of the registration area.

LC Hosted Dinner

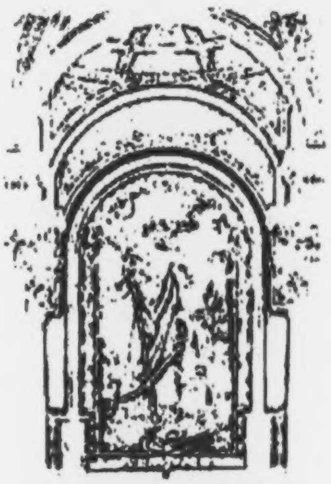
Great Hall
Thursday, November 16, 2000
Thomas Jefferson Building
1st Street, SE map
To access the Great Hall, please present Conference Badge
First Floor Entrance - Clear Security
Proceed to self service coat checking and or registration/information table
6:30 - 7:30 p.m. Reception/Exhibit-Mezzanine
7:30 - 8:45 p.m. Dinner

Complimentary Box Lunches

Friday, November 17, 2000
11:45 a.m.
You may pick-up your gourmet box lunches on the 6th floor in the registration area. The Dining Room A and the Mumford Room are

The LIBRARY *of* CONGRESS CATALOGING

Cataloging Frequently Asked Questions (FAQ)
About the Cataloging Directorate - Search Online Catalogs
Programs & Services - Publications & Newsletters
Reports & Proceedings - Related Products & Services



Mosaic by Elihu Vedder
Visitor's Gallery

NEW! New in Cataloging NEW!
LC Report from the 2001 MLA/MOUG meeting in
PDF format or WordPerfect8 format
Fiscal Year 2000 Annual Report for the Cataloging Directorate
Bicentennial Conference on Bibliographic Control
in the New Millenium along with
cybercasts of the conference proceedings
Cybercast video of a staff Cataloging Forum presentation from April 2000 on
CORC, the BEOnline+ and BECites+ projects, and CORC pathfinders at LC
LC Pinyin Conversion Project

The Cataloging Directorate's mission is to provide innovative and effective bibliographic control of the Library's collections and leadership to the library and information communities in the development of cataloging theory and practice.

About the Cataloging Directorate

- How to Contact the Library Regarding Cataloging Questions
- Cataloging Directorate Management Team
- Modes of Cataloging Employed in the Cataloging Directorate
- Archive of older Cataloging Directorate memos, messages, and announcements

Library of Congress Cataloging Policy and Practice

Cataloging Policy and Support Office (CPSO) Home Page

Available at this site: News on changes or proposed changes in cataloging policy and practice at the Library, as well as cataloging tools and documentation, information on name authorities, subject headings (including weekly lists of new and changed subject headings), and the Library of Congress Classification.

Search the Library's Online Catalogs

Forms-based Search of the Library of Congress Files

Search the LC database or other library catalogs by using a WWW search form and Z39.50 technology.

Other Public Access to the Library's Online Catalog

Search catalog records for books, microforms, serials, cartographic items, music, visual materials (films, filmstrips, posters, prints, photographs), computer files, manuscripts, thesauri of names and subject terms used in cataloging records, etc.

Cataloging Programs and Services

Program for Cooperative Cataloging (PCC)

PCC is an international cooperative effort aimed at expanding access to library collections by providing useful, timely, and cost-effective cataloging that meets mutually-accepted standards of libraries around the world.

National Union Catalog of Manuscript Collections (NUCMC)

NUCMC is a free-of-charge cooperative cataloging program operated by the Library of Congress. NUCMC catalogers create MARC bibliographic records to describe collections held by participants, and establish pertinent name and subject authority headings.

Cooperative Program for Serials Cataloging (CONSER)

CONSER is a cooperative online serials cataloging program, a source of high quality bibliographic records and documentation for the cataloging of serials.

Serial Record Division

The Serial Record Division has broad responsibility for the processing and bibliographic control of the Library's serials. The Serial Record is the nation's largest serials file, including approximately 900,000 serial entries. Of these, 150,000 titles are estimated to be current receipts. Over 20,000 new entries are added to the file each year.

Information for Publishers

Contact information for obtaining an International Standard Book Number (ISBN), the U.S International Standard Serials Number (ISSN) Center Home Page, and descriptions of the Preassigned Card Number and Cataloging in Publication programs.

The Bibliographic Enrichment Advisory Team (BEAT)

BEAT is charged with the development and implementation of initiatives to develop tools to aid catalogers, reference specialists, and searchers in creating and locating information, seeks to enrich the content of Library of Congress bibliographic records as well as improve access to the data the records contain, and conducts research and development in areas that can contribute to furthering these efforts.

Publications and Newsletters

LC Cataloging Newsline (LCCN)

LC Cataloging Newsline is the electronic journal of the Cataloging Directorate. Articles are selected on the basis of interest to the broader cataloging community.
LCCN back issues

CONSERline

CONSERline is the electronic journal of the CONSER program, the Cooperative Online Serials program

administered at the Library of Congress.
- CONSERline Home Page

Reports and Conference Proceedings

Fiscal Year 1999 Annual Report for the Cataloging Directorate

Paper presented to the Authorities Subcommittee, Bibliographic Control Committee of the Music Library Association on Machine-Derived Name Authority Records

Proceedings

- *Seminar on Cataloging Digital Documents* (October 12-14, 1994)
- *Organizing the Global Digital Library Conference* (December 11, 1995)
- *Organizing the Global Digital Library II and Naming Conventions* (May 21-22, 1996)
  ASCII version on LC MARVEL

Other Reports and Papers via the LC MARVEL Gopher

Related Products and Services Outside the Cataloging Directorate

Cataloging Distribution Service (CDS)

CDS develops and markets products and services which provide access to Library of Congress resources. You can order CDS products from the Complete Catalog of Bibliographic Products and Services which is searchable by title or subject.

MARC Home Page

The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.

Go to:

- Top of Page
- Using the Library: Collections and Services
- Search or Browse the Library of Congress Web Site
- Library of Congress Home Page

Library of Congress
Comments: lcweb@loc.gov (05/13/99)

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Organizing Team

John D. Byrum, Jr. is chief of the Regional and Cooperative Cataloging Division. He is responsible for providing leadership and management of activities to accomplish planning and preparations for as well as implementation of the Bicentennial Conference.

BEST COPY AVAILABLE

Cataloging Directorate Home Page

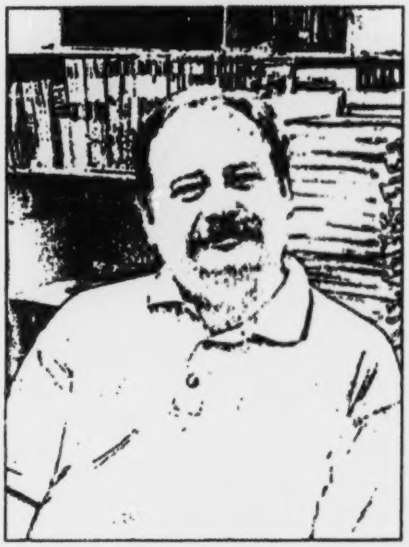
Library of Congress Home Page

Ann Sandberg-Fox is an independent cataloging consultant and trainer, with offices located in Fairfax, VT. She has been appointed to serve as Conference Consultant and is responsible for providing expert advice regarding the content and format of the conference and for assisting the Conference Management Team on a number of fronts.

Cornelia Owens Goode, Program Specialist, Regional and Cooperative Cataloging Division, is serving as Conference Administrative Coordinator. Ms. Goode is responsible for lodging accommodations, administrative and logistical support.

Bruce Chr. Johnson, Senior Library Information System's Specialist and Team Leader, Cataloger's Desktop/Classification Plus Development Team, Cataloging Distribution Service, is serving as Conference Budget and Publications Officer. He is responsible for overseeing conference expenses, initiating fundraising efforts, and handling arrangements for publication of the proceedings.

David Williamson, Cataloging Automation Specialist in the Cataloging Directorate, provides automation planning and support for the conference. He serves as webmaster for the conference web pages, listowner for the discussion list to be used to discuss conference topics, and is coordinating the effort to cybercast speakers.



Judy Mansfield, Chief of the Arts and Sciences Cataloging Division, is responsible for the conference topical discussion groups, including announcing the groups to the Conference attendees and enrolling the attendees in the various groups in consultation with Ann Sandberg-Fox to achieve a balance of numbers and expertise.

Susan R. Morris, Assistant to the Director for Cataloging, coordinated note taking for the Conference plenary sessions and assisted with preparing summaries of the Topical Discussion Groups' recommendations. She is now handling post-Conference publicity.

Library of Congress
January 30, 2001
Comments: lcweb@loc.gov

What's new

Cybercast information

Candid photos from the conference

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Topical Discussion Group Recommendations

The Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium was both provocative and extremely productive. The conference, held on Nov. 15-17, 2000, featured eleven breakout sessions, or Topical Discussion Groups, that addressed major challenges facing catalogers and their allies in the vendor and publisher communities. Each Topical Discussion Group (TDG) presented a set of recommendations to the final plenary session of the Conference, and the recommendations were later circulated for additional input from all Conference participants. The summaries of the TDGs' work, revised to take the Conference participants' comments into account, are now available for all members of the library and information communities to read.

You can view the recommendations in PDF format by clicking on the names of the Topical Discussion Groups below:

1. Choosing Electronic Resources: What Is a Valuable Web Resource?
2. What Are The Continuing Education Needs of Professional Catalogers?
3a. What Near-Term Cooperative Partnerships Should Libraries Explore in the Digital World?
3b. What Long-Term Cooperative Partnerships Should Libraries Explore in the Digital World?
4a. Multiple Versions (Originally titled How Can AACR2 Become More Responsive to Cataloging Networked Resources on the Web in the Near-Term?)
4b. How Can AACR2 Become More Responsive to Cataloging Networked Resources on the Web?
5. What Can the Library Community Offer in Support of Semantic Interoperability?
6. What Automated Tools Could Assist Libraries to Meet the Information Needs of Their Users?

7. What Steps Can The Library Take to Achieve Integrated Access to the Catalog and Other Discovery Tools?
8. How Can Libraries Participate More Actively in the Development of Metadata Standards?
9. How Can Catalogers and Metadata Providers Ensure that Resource Descriptions Meet Reference Needs?

The Library welcomes comments on the TDG recommendations from interested readers at any time. Please email comments to John Byrum, chair of the Conference Organizing Team, at jbyr@loc.gov. Library management and staff are now developing a plan for addressing the many recommendations, in order to determine which are feasible to adopt in the short and long terms. Some of the recommendations would require the Library to seek additional funding or other resources in order to implement them, and some recommendations are in conflict with each other. The Library cannot guarantee that all the recommendations will be carried out, but it does assure all readers of the Conference Web site that their comments will be considered, as plans evolve.

– Beacher Wiggins, Director for Cataloging, Library of Congress

- Conference Organizing Team
- Cataloging Directorate Home Page
- Library of Congress Home Page

Library of Congress
January 04, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Cybercasts Now Available

In order to view this cybercast, you will need either Netscape 4.x or higher or Internet Explorer 5. You will also need Real Player version 8 or higher. If any of your components are out of date, you should get an error message and the opportunity to upgrade those components. You will also need to have a good internet connection for proper viewing, such as an institutional internet connection. Dial-up connections with a 56.6K modem will prove unsatisfactory.

If you get a message about a missing plugin that has no available update, you should go to real.com and get the latest free version of Real Player to install on your machine. That should clear up any problems.

If network traffic is heavy, you will see the synchronization between the speaker and his/her voice start to slip and the graphics may be slow to change. This does not cause any problems, it is just a sign that network traffic is heavy.

Day 1 Speakers:

Introductory remarks by Beacher Wiggins, Director for Cataloging (11:17)

Introductory remarks by John Byrum, Chief, Regional and Cooperative Cataloging Division and conference organizer (6:01)

Keynote address "From Card Catalogues to WebPACS: Celebrating Cataloging in the 20th Century" by Michael Gorman (17:35)

Topic 1: The Library Catalog and the Web

Cataloging Directorate Home Page

Library of Congress Home Page

The Catalog as Portal to the Internet by Sarah Thomas (19:35)

The Library Catalogue in a Networked Environment by Tom Delsey (22:02)

International Metadata Initiatives: Lessons in Bibliographic Control by Priscilla Caplan (19:01)

Comments on Thomas paper by Brian Schottlaender (12:25)

Comments on Delsey paper by Jennifer Trant (12:45)

Comments on Caplan paper by Robin Wendler (9:50)

Q & A session relating to topic 1 speakers (14:40)

Topic 2: Assessing Current Library Standards for Bibliographic Control and Web Access

Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources by Matthew Beacom (20:21)

Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources by Lois Mai Chan(19:41)

Comments on Beacom paper by Glenn Patton (12:41)

Comments on Chan paper by Diane Vizine-Goetz (11:50)

Resource Discovery Using Z39.50: Promise and Reality by William E. Moen (22:05)

Authority Control on the Web by Barbara Tillett (18:37)

Q & A session relating to topic 2 speakers (14:50)

Topic 3: Future Directions

AACR2 and Its Place in the Digital World: Near-term Revisions and Long-term Direction by Ann Huthwaite (18:50)

Extending MARC to Meet New Challenges in Bibliographic Control of the Web by Sally McCallum (20:54)

BEST COPY AVAILABLE

Comments on Huthwaite paper by Lynne Howarth (16:15)

Comments on McCallum paper by Paul Weiss (9:45)

Day 2 Speakers:

Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog? by Carl Lagoze (20:55)

Descriptive Resource Needs from the Reference Perspective by Carolyn Larson and Linda Arret (21:12)

Q & A session relating to topic 3 speakers (20:50)

Topic 4: Experimentation

A Comparison of Web Resource Access Experiments: Planning for the New Millennium by Jane Greenberg (15:20)

Redesign of Library Workflows: Experimental Models for Electronic Resource Description by Karen Calhoun (20:52)

Q & A session relating to topic 4 speakers (15:04)

Topic 5: Exploring Partnerships

Exploring Partnerships: Librarians, Producers and Vendors: What Do Librarians Need? by Michael Kaplan (20:10)

Comments on Kaplan paper by Lynn Connaway (12:35)

Comments on Kaplan paper by Jeff Calcagno (12:50)

Comments on Kaplan paper by Amira Aaron (14:16)

Partnerships to Mine Unexploited Sources of Metadata by Regina Reynolds (18:55)

Q & A session relating to topic 5 speakers (5:37)

Day 3 Topical Discussion Group recommendations

TDG1: Choosing Electronic Resources: What Is a Valuable Web Resource? (9:25)

TDG2: What Are The Continuing Education Needs of Professional Catalogers? (10:41)

TDG3A: What Near-Term Cooperative Partnerships Should Libraries Explore in the Digital World? (11:09)

TDG3B: What Long-Term Cooperative Partnerships Should Libraries Explore in the Digital World? (13:41)

TDG4A: How Can *AACR2* Become More Responsive to Cataloging Networked Resources on the Web in the Near-Term? (13:57)

TDG4B: How Can AACR2 Become More Responsive to Cataloging Networked Resources on the Web in the Long-Term? (15:53)

TDG5: What Can the Library Community Offer in Support of Semantic Interoperability? (5:49)

TDG6: What Automated Tools Could Assist Libraries to Meet the Information Needs of Their Users? (13:39)

TDG7: What Steps Can The Library Take to Achieve Integrated Access to the Catalog and Other Discovery Tools? (12:14)

TDG8: How Can Libraries Participate More Actively in the Development of Metadata Standards? (12:14)

TDG9: How Can Catalogers and Metadata Providers Ensure that Resource Descriptions Meet Reference Needs? (8:09)

Wrap-up and closing remarks by Beacher Wiggins, Director for Cataloging (9:58)

Cataloging Directorate test cybercast video on CORC, BEOnline+, BECites+, and CORC Pathfinders

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

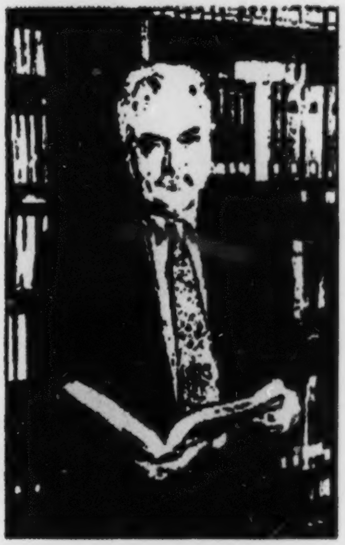
Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate



Michael Gorman

Dean of Library Services
California State University
Madden Library
5200 N. Barton
Fresno, CA 93740-8014

Keynote Address: From Card Catalogues to WebPACs: Celebrating Cataloguing in the 20th Century

About the presenter:

Michael Gorman is Dean of Library Services at the Henry Madden Library, California State University, Fresno. From 1977 to 1988 he worked at the University of Illinois, Urbana, Library as, successively, Director of Technical Services, Director of General Services, and Acting University Librarian. From 1966 to 1977 he was, successively, Head of Cataloguing at the British national bibliography, a member of the British Library Planning Secretariat, and Head of the Office of Bibliographic Standards in the British Library. He has taught at library schools in his native Britain and in the United States--most recently as Visiting Professor at the University of California, Berkeley, School of Library and Information Science (summer sessions).

He is the first editor of the *Anglo-American Cataloguing Rules*, second edition (1978) and of the revision of that work (1988). He is the author of *The Concise AACR2* (1989); editor of, and contributor to, *Technical Services Today and Tomorrow*, 2nd edition (1998); and editor of *Convergence* (proceedings of 2nd National LITA Conference), and *Californien*, both published in 1991. *Future Libraries: Dreams, Madness, and Reality*, co-written with Walt Crawford, was honored with the 1997

Cataloging Directorate Home Page

Library of Congress Home Page

Blackwell's Scholarship Award. His most recent book, published by ALA in 1997, is titled *Our Singular Strengths: Meditations for Librarians*. Mr. Gorman is the author of more than 100 articles in professional and scholarly journals. He has contributed chapters to a number of books and is the author or editor of other books and monographs. He has given numerous presentations at international, national, and state conferences.

Michael Gorman is a fellow of the [British] Library Association, the 1979 recipient of the Margaret Mann Citation, the 1992 recipient of the Melvil Dewey Medal, and the 1997 recipient of Blackwell's Scholarship Award.

Full text of paper is available

Summary: This keynote address recounts the many important accomplishments and advancements in cataloguing theory and practice which have occurred between 1900 and 1999, and provides a backdrop for the papers and discussions which follow. The address also serves as an upbeat reminder of all the progress that has been made and, we hope, will inspire conference participants to tackle the challenges of networked resources and the Web with enthusiasm and resolve.

Library of Congress
May 9, 2000
Comments: lcweb@loc.gov

From Card Catalogues to WebPACS:

Celebrating Cataloguing in the 20th Century
a talk given at the
Library of Congress Bicentennial Conference on Bibliographic Control
for the New Millennium Washington, D.C., November 15th 2000

Michael Gorman
Dean of Library Services
California State University, Fresno

Final version

> I dreamed I saw Joe Hill last night,
> Alive as you or me:
> Said I, but Joe you're ten years dead;
> I never died said he.
> I never died said he.
> And standing there as big as life
> A-smiling with his eyes.
> Said Joe, what they forgot to kill
> Went on to organize,
> Went on to organize.
> *The Ballad of Joe Hill*
> *by Alfred Hayes and Earl Robinson (1925)*

Introduction

The story of cataloguing in the 20th century is the story of two structures. The first is that of codes and standards—from the 4th edition of Cutter's rules in 1904 through the Red Book, ISBDs, MARC, and AACR2. The second is that of the means by which catalogue records are communicated—from the book catalogues and cards of the turn of the century through microfiches, online catalogues, and Web-based catalogues. Both are a story of onward and upwards, but both are threatened by the bizarre millenarianism of "the end of history" crowd. In their view, forms of catalogue are irrelevant since all forms of human communication will be swept away in favor of digital communication (and those digital documents will, mysteriously, catalogue themselves). Also, to them cataloguing standards are unimportant since they believe they do not apply to Web sites and the rest. Such views are not only wrong but also noxious because, though masquerading as progressive, they are impeding progress. Digital communication is an important development, but it is not a unique and obliterating development, as any historian of communication will attest. Contemporary cataloguing standards not only can be used for digital resources, but are also greatly superior to the generally ill-considered proposals that are advanced as answers to the wrong question. "How should we catalogue electronic resources?" is not an important question. "Which electronic resources should we catalogue and how shall we preserve them?" is. The reason is that effective cataloguing involves controlled vocabularies and adherence to the standards that have evolved in the past 100 years.

What, fundamentally, is the topic of this conference? It is the idea of describing assemblages of recorded knowledge and information in terms of their titles, editions, issuers, date, extent, etc.; of adding formalized names and titles to those descriptions that allow library users to retrieve and collocate those descriptions; and relating them to a location—physical or in cyberspace. That is it, but one might as well describe chess as a game in which 32 pieces of wood or plastic are moved on a cardboard checkerboard according to prescribed rules. True, but scarcely an explanation of the fascination of cataloguing or chess or of the unlimited permutations and problems to be solved. Charles Ammi Cutter knew this, when he wrote, in 1904, of cataloguing's "... difficulties and discussions which have furnished an innocent pleasure to so many ..."[1] Cutter made that remark in the context of his idea, widely shared, that the advent of the LC printed catalogue would resolve all those difficulties and discussions and that the energy hitherto put into cataloguing would be diverted into "other parts of the service—the children's room and the information desk, perhaps."[2] We all know that his idea did not exactly come true and cataloguing in the 20th century turned out to be full of difficulties and discussions.

Just as in the early 1900s, there is a tendency today to belittle the importance of descriptive cataloguing, even by people, who should know better (I shall deal with them later.) The difference is that Cutter and others thought that cataloguing had been perfected, whereas the naysayers today believe that cataloguing is irrelevant.

Issues in 1901

Let us begin by looking back 100 years to the cataloguing issues that preoccupied our long-dead colleagues in 1901. C.W. Andrews' paper at a meeting of the Illinois Library Association in February 1901 dealt with issues raised by the emerging cooperative cataloguing system. He welcomed the economy, fullness, uniformity and legibility of the printed cards. In a back to the future moment, Mr. Andrews opined "... the effect of this plan will not be to deprive catalogers of their work, but to substitute the intellectual for the mechanical ..."[3]—a point that I found myself making over and over again, also in Illinois, some 75 years later. The Advisory Committee on Cataloging, formed by the ALA Publishing Board, met in Atlantic City in March 1901, Cutter and other luminaries in attendance.[4] They considered the typography of printed catalogue cards, discussed and rejected a proposal that the contents note be given after the title (a far more radical change in order of bibliographic data that anything envisaged by MARC or the Dublin Core), and stated that full names were more important in headings for English authors and in large libraries. The committee agreed that headings should be those to be found "...where the average person using a library is apt to look ..." The report gives this as a throwaway line, seemingly without consciousness that that the great fault line between "correct" headings and "sought" headings was being established by this reference to the needs of the "average person." They then moved on to wrestle with corporate entries—generally agreeing with Cutter's rules (then in their 3rd edition), which stipulated that some corporate bodies were to be entered under name and some under place. This innocent seeming decision was to bedevil cataloguing and cataloguers for the next 77 years. They decided that there should be a more extensive use of birth and death dates in headings. The most divisive issue was that of measuring the size of books. The Committee could not choose which of three proposed methods (letter symbols, fold symbols, or exact size in centimeters) should be used, but did decide to write minority reports on each. The report states: "The committee has been impressed with the practical agreement of its members on cataloging rules, upon the willingness to yield on inessential points, and upon the idea that the catalog should be made for the user, not for the cataloger."

Before we succumb to the temptation to see Cutter and his colleagues as quaintly old-fashioned, let us remember that they were far more advanced in the standardization of cataloguing the materials of their age, and in cooperating, than we are in dealing with the electronic documents of our day. In fact, the parallel with our time is the situation in the late 18th century when the French revolutionaries hit on the idea of using the blank backs of playing cards to record the holdings of the aristocratic libraries taken over in the name of the people. (On reflection, that is a far more organized and coherent scheme than anything we have done for electronic documents to date.) The late 18th century was a time of chaos, to which a few brave souls tried to bring order, one small step at a time. Ours is a culture in chaos—a time of beleaguered learning and of threats to the records of humankind. We too need a few brave souls, and should applaud those who try to bring real cataloguing to bear, while defying those who want to capitulate to the fecklessness that disregards standards and bibliographic control, on the irrelevant and dubious grounds that electronic documents are transcendent and transformational.

1908 code

In 1908, committees of the ALA and the [British] Library Association published *Catalog rules: author and title entries*[5] in two editions, thus setting an unfortunate precedent that lasted until 1968. The committees were unable to agree on all rules, both between the US and the UK, and between themselves and LC practice. LC was more robust in those days and, rather than issuing *Rule interpretations* that contradict the rules, you will find in the 1908 rules flat statements of LC practice that differs from the rules in matters great and small. Thus, a British or American cataloguer would, for the next 40 years, have to choose, in many instances, between British rules, American rules, and LC rules. In North America, in which the LC card was to dominate cataloguing for at least the next 80 years, there was a strong tendency to follow LC practice and to hell with the rules to which LC practice was an alternative.

The 1908 code was dominated by cases not conditions and principles. This arose because of the bilateral nature of its origins and construction. Nineteenth century codes were almost all the product of single individuals (Panizzi, Cutter, Jewett, etc.). Despite the fact that Cutter was active on the American committee until his death in 1904, his was but one voice among many, often eminent others. Lacking a guiding hand and a single set of unified principles, this, the first of the committee codes that have lumbered through the 20th century, was an assemblage of the best practices of Anglophone libraries. It was inevitable that such a code would be based on cases and ever more minute distinctions between cases. The latter reached its *reductio ad absurdum* in the full page devoted to the rule on Exploring expeditions, with its two subrules, the second of which has 6 sub-sub-rules. The 1908 code set in train a period of code making that was to lead, inevitably, to calls for reform from Andrew Osborn and others.

Vatican code

There was an interesting statement of American cataloguing practice in the inter-war years. The Vatican Library published a code of rules in 1931[6] that was later stated to be "... the most complete statement of American cataloging practice."[7] The Vatican code was notable for a number of reasons. It included rules on name and title entry, description, subject headings, and filing—the only code since Cutter to do so. Had it not been for the Second World War, it is quite possible that work on the Vatican code would have taken the place of the work that led to the abortive 1941 draft rules and the unmitigated disaster of the 1949 ("Red Book") rules. Perhaps, however, it was not just the course of cataloguing that was changed irrevocably by WWII.

1941 and Osborn

The tenuous connection between North American and British cataloguing committees broke entirely with the publication of the 1941 *ALA cataloging rules, preliminary American second edition*[8] (the British being largely occupied with other matters in 1941). The main contribution of the 1941 draft rules to cataloguing history was the reaction it provoked in the Australian librarian Andrew Osborn. His *The crisis in cataloging*[9] called, in essence, for fewer, simpler rules based on principles and ignoring non-essentials. He also called for codes that allowed cataloguers to use their judgment based on experience and, again, on principles. Perhaps naively, he thought that such cataloguers would win more respect from library administrators. Osborn's important article appeared to have faded into oblivion as the attention of the United States turned to World War II, which the Americans entered a little over a month after The crisis was published. After the war, the ALA/LC cataloguing committees carried on their work as if Osborn had never spoken. The result was "The Red Book" and "The Green Book"[10] of 1949.

1949: ALA and LC

Almost all you need to know about the Red Book (The *ALA cataloging rules for author and title entries*[11]) is summed up in the fact that one rule, 116A(3), is devoted to *and only* to the Basilian Monastery at Mount Sinai. This is the logical inevitable result of piling case upon case and splitting ever thinner hairs, all the while ignoring the principles or even the need for principles and ignoring the needs of catalogue users for clarity and consistency. After 1949, there were only two possible directions. Reform or progress toward a code that consisted of nothing but cases applying to tiny numbers of documents. Thank the Lord and Lubetzky, we embarked on the road—the long, twisty, and obstacle-ridden road—to reform.

Lubetzky

Seymour Lubetzky, employed at the time by the Library of Congress, was the most prominent critic of the 1949 rules.[12] His seminal work *Cataloging rules and principles*[13] was subtitled "a critique of the ALA rules for entry," but might, like a latter day John Knox, have been entitled "A blast against the monstrous regiment of cataloguing rules." With the simplicity of genius, Lubetzky stepped away from the trees of exploring expeditions and Basilian monasteries and saw the forest of the cataloguing code in asking his famous question "Is this rule necessary?" Further, he asked of every rule is it consistent with principles and is it properly related to other rules. These questions gave rise to a draft code[14] that was as spare and coherent as the 1949 rules were sprawling and incoherent. The Lubetzkyan revolution spilled over into the "Paris principles"[15] which were thought, at the time, to be the framework for a universal cataloguing code that would revolutionize international bibliographic cooperation. Alas, reality intervened in the unholy alliance of traditionalist cataloguers (some not a million miles away from the Library of Congress) and library administrators (the forces of darkness then and in the War of AACR2) that caused Seymour Lubetzky to be replaced as editor of the code that was aborning—the code that was intended to unite North American and British cataloguing and usher in the Lubetzkyan age of global cooperation.

1968: two codes

The sad fact is that the code that resulted from this reactionary tide—the first AACR[16]--was not only a major fudge betraying Lubetzky's ideas in many instances but also could not even reconcile British and American practice (the prophet Lubetzky was honored far more in the UK than on his native heath). Also, and crucially, AACR failed to deal adequately with what we used to call "non-books." The first AACR did have many strengths, and was a great improvement over its predecessors, but, ultimately, it represented a failure of nerve that has consequences to this very day. I cannot now give details of the good and bad rules it contained, but wish merely to emphasize that its failure was truly historic in that it failed to live up to its time—an era in which Universal Bibliographic Control became more than a dream. In that era, a unified English-language cataloguing code based on coherent principles would have saved us from a couple of decades of squabbles and confusion.

1968: MARC and ISBD

So, what happened? MARC and ISBD happened—both in the late 60s and early 70s. It should be unnecessary to point out that MARC is not a cataloguing content standard—it is a framework standard to which cataloguing content has to be added. I mention this only because I have heard and read so many of the metadata boys talking about "MARC cataloguing," a shockingly ignorant phrase that betrays the fact that the very nature of cataloguing is not understood in such circles. In MARC, the electronic version of the catalogue card as a carrier of bibliographic data, we had a means of communicating bibliographic data between nations and continents—a means that would be utterly ineffective without universally agreed standards for the content of bibliographic records. The International Standard Bibliographic Description (ISBD) provided a complementary framework to MARC, and some guidance on the international standardization of descriptive content. The stage was set for a totally new code for a new bibliographic world.

1978: AACR2

It is quite unnecessary to rehearse all the events and circumstances that led to the publication, in 1978, of AACR2 beyond saying that was modified and resisted because of pressure from the same reactionary elements that stalled the Lubetzkyan revolution. AACR2's very name is a fraud. This was by no means a second edition of the 1968 code, but a new code that should have had its own name, something that would have spared us much subsequent grief. Concern about the amount of change the new code would mandate mingled with a witches' brew of nonsense about the ISBD being "unnatural" and "foreign" (nonsense that is being repeated to this day) and pressure from special classes of cataloguer to produce a code that is encrusted with the unnecessary elaborations about which both Osborn and Lubetzky complained. There is a major difference, however. AACR2 *is* barnacled with unnecessary rules, but has a core of Lubetzkyan truth in its rules on entry and heading. Moreover, the descriptive rules, based as they are on the ISBD, have a coherent structure that is fair in its treatment of all media, and capable of incorporating any new media. These strengths were shown in the publication of AACR2R--a publication that caused none of the excitement of AACR2 but quietly demonstrated the strength of the underlying structure of the 1978 code.

In the 19th century, we had codes that were the product of individuals, for most of the 20th century, we had codes that were the product of committees. Then, in 1998, we reached the next step—a code that was the product of nobody at all! The 1998 publication (I scorn to call it an edition or revision) of AACR2 consisted of the 1988 revision with agreed amendments slipped in without regard for their implications elsewhere in the text. I suppose a virtual code was inevitable. Perhaps a robot called Hal will produce the next? Whatever that future may be, the 1998 publication is indicative of the parlous state of cataloguing codes at the close of this century—no battles, no excitement, but no progress either.

OPACs/WebPACs

Perhaps metadata is the reason why progress on refining the cataloguing code appears to be at a standstill? Now what? It used to be AACR3? Now, the chatter is that all that matters are electronic documents and "traditional" standards cannot deal with electronic documents. Neither happens to be true, but that has not stopped the most abundant supply of hot air (from proponents of metadata) since the War of AACR2 and the great "ISBD is a foreign plot" debate.

Metadata, as I said previously, is a subset of the MARC record without the instructions on standardization of content necessary to create a bibliographic control system. Amusingly, the Dublin Core (like MARC) is based on the catalogue card in that it preserves the main entry (see the placement of the principal creator and subsidiary names). Metadata (a fancy name for an inferior form of cataloguing) also ignores the central question—what should be catalogued? No one is interested in taming all the vast wasteland of the Internet because most of that vast wasteland is worthless. The task is to identify the oases and to apply some version of real cataloguing to them. My belief is that a modicum of common sense is beginning to dawn. A recent discussion in the California State University libraries centered on the fact that many of our patrons were unaware of the many journals that are available online (at some considerable expense to the taxpayer). My mind had wandered to some more agreeable topic (utterly unconnected with libraries) but I was brought back abruptly when a consensus emerged that use of these journals might well be increased if they were … wait for it, listed individually in the catalogue!

So, here we are at the end of a century of erratic progress in cataloguing, facing the possibility that we will settle for tenth best, a weird amalgam of free text searching and unstandardized, uncontrolled, ersatz cataloguing masquerading as a branch of information "science." However, there is another possibility—that we will incorporate electronic documents into Universal Bibliographic Control, as we have incorporated all other forms of human communication and thus usher in another golden age of cataloguing that supports our unique task as librarians—the preservation and onward transmission of the human record.

Thank you.

Notes:

1. Cutter, C.A, Rules for a printed dictionary catalogue. 4th ed. Washington, DC: GPO, 1908. Introduction.
2. *Ibid.*
3. Roper, Eleanor. Illinois State Library Association. *Library journal,* v.26, no. 3 (March 1901) p. 146.
4. Kroeger, Alice B. Advisory Committee on Cataloging Rules. *Library journal,* v. 26, no.4 (April 1901) pp. 211-212
5. Catalog[u]ing rules: author and title entries. North American ed. Chicago: ALA, 1908. English ed. London: LA, 1908.
6. *Norme per il catalogo degli stampati* of the Biblioteca apostolica vaticana were published in 1931 (2nd ed. 1939). Thomas J. Shanahan translated *Selected rules for author and title entries* from the *Norme* and it was issued (though not, I think, "published") by St. Paul Seminary (St. Paul, Minn.) in 1931. A full translation (of the 2nd Italian ed.) was prepared and circulated in 1939 but not published until 1948 (Chicago:ALA).
7. *Norme (op.cit.)* Foreword to the English translation of the 2nd Italian edition.
8. ALA cataloging rules. Preliminary American second ed. Chicago: ALA, 1941.
9. Osborne, Andrew. The crisis in cataloging. *Library quarterly.* v. 11, no.4 (October 1941) pp.393-411.
10. Rules for descriptive cataloging in the Library of Congress (adopted by the American Library Association). Washington, DC: LC, 1949.
11. ALA cataloguing rules for author and title entries. Second edition/ edited by Clara Beetle. Chicago: ALA, 1949.
12. Gorman, Michael. Seymour Lubetzky, man of principles. *In* The future of cataloging: insights from the Lubetzky Symposium. Chicago: ALA, 2000. pp12-21.
13. Lubetzky, Seymour. Cataloging rules and principles. Washington: LC, 1953.
14. Lubetzky, Seymour. Code of cataloging rules: an unfinished draft. Chicago: ALA, 1960.
15. International Conference on Cataloguing Principles, *Paris, October 1961.* Statement of principles.

Sevenoaks, Kent, England: IFLA, 1966.
16. Anglo-American catalog[u]ing rules. North American text. Chicago:ALA, 1968. British text. London: Library Association, 1968.
17. Anglo-American cataloguing rules, second edition. Chicago: ALA, 1978.
18. Anglo-American cataloguing rules, second edition, revised. Chicago: ALA, 1988

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Clifford Lynch

Executive Director,
Coalition for Networked Information

The New Context for Bibliographic Control In the New Millennium

About the presenter:

Clifford Lynch has been the Director of the Coalition for Networked Information (CNI) since July 1997. CNI, jointly sponsored by the Association of Research Libraries and Educause, includes about 200 member organizations concerned with the use of information technology and networked information to enhance scholarship and intellectual productivity. Prior to joining CNI, Lynch spent 18 years at the University of California Office of the President, the last 10 as Director of Library Automation. Lynch, who holds a Ph.D. in Computer Science from the University of California, Berkeley, is an adjunct professor at Berkeley's School of Information Management and Systems. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science and the National Information Standards Organization. Lynch currently serves on the Internet 2 Applications Council; he was a member of the National Research Council committee that recently published The Digital Dilemma: Intellectual Property in the Information Infrastructure, and now serves on the NRC's committee on Broadband Last-Mile Technology.

Full text of comments is available

Summary:

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

Supporting the identification of works of interest is not the only purpose of bibliographic control, but it is certainly one of the most important and most widely relied-upon. In this paper I will consider the ways in which information finding is changing in a world of digital information and associated search systems, with particular focus on methods of locating information that are distinct from, but complementary to, established practices of bibliographic description. A full understanding of these developments is essential in re-thinking bibliographic control in the new millennium, because they fundamentally change the roles and importance of bibliographic metadata in information discovery processes.

There are three major approaches to finding information: through bibliographic surrogates, that represent an intellectual description of aspects and attributes of a work; through computational, content-based techniques that compare queries to parts of the actual works themselves; and through social processes that consider works in relationship to the user and his or her characteristics and history, to other works, and also to the behavior of other communities of users.

The first approach is familiar, and forms the basis of catalogs and abstracting and indexing, and more recently online catalogs and similar systems. The third approach is also familiar, in the form of book reviews, citation indexes, and suggestions from colleagues, but is now seeing a great creative expansion in the digital world, with its ability to create and aggregate world-wide communities of interest and to track the behavior of users. The second is fundamentally new in the digital world, where techniques based on full text searching form the basis of today's web search engines. We need to recognize that in the new millennium, for digital materials, high quality content-based computational techniques will be an inexpensive, ubiquitous, and rapidly-available default means of searching, and that powerful socially based approaches will also be widely available at little cost.

This leaves us with a number of challenges for bibliographic description in the new millennium. What are the unique contributions of approaches based on human intellectual analysis? When are they justified, and on what basis? Can we devise a spectrum of bibliographic approaches, with an accompanying spectrum of costs, to complement the content-based and socially-based approaches? How do we most effectively fuse the three approaches into information discovery systems that are truly responsive to user needs?

There is an additional set of questions that need to be considered as part of mapping the context for the new bibliographic control.

First, we know that bibliographic control is not just about rules and practices. It also depends upon a rich and complex infrastructure of authority files and classification structures. Indeed, the other approaches also use infrastructure - for example, lexicons, dictionaries, gazetteers and similar tools for content-oriented computational techniques, and methods to manage identity, authenticity, and reputation in the case of socially-based systems. It will be important to determine how much of this infrastructure can be shared, and leveraged, among the three approaches, and what the practitioners of each approach can do to enhance this.

Second, we must recognize the democratizing and empowering character of the networked information environment; just as anyone can become a distributor of information with a global reach, anyone can become a describer of information. Metadata itself is information, and we need to be able to decide when we choose to trust it; thus many of the same tools and techniques that have become relevant to the socially based discovery of information in the digital world will also become applicable in the production and use of bibliographic metadata - the linkage of metadata to identities through digital signatures, the management of identities through public key infrastructure, and the manipulation of reputation related to these identities. Thus we have a specific challenge in understanding how to connect and apply the infrastructure that is being driven by the social techniques - and indeed by much broader developments in the networked environment, such as electronic commerce - to bibliographic control.

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

The New Context for Bibliographic Control In the New Millennium

Clifford Lynch

Final version

This text is based on a dinner speech given in the Great Hall of the Jefferson Building of the Library of Congress on the evening of November 16, 2000 as part of the Bibliographic Control for the New Millennium conference.

The broad reading public, from scholars to students, researchers to recreational readers, wants and needs to find works of relevance to their interests. Enabling the identification of such works meeting these needs is not the only purpose of bibliographic control, but it is certainly one of the most important and most widely relied-upon. It is the most visible, and, to a great extent, the reason why there is support for the very substantial ongoing investment in bibliographic control. But the practices of information finding are changing in a world of digital information and computer-based search systems. Within the library community we have placed great emphasis on the impact of on-line public access catalogs and abstracting and indexing databases designed to be searched by the general public rather than specially trained intermediaries for several reasons. These systems arrived early; they have been deployed for about two decades on a reasonable scale and are now deployed almost ubiquitously. They are effective and well-received by their users; they represent a very significant improvement in the quality of access to library collections. And the library community is, at some fundamental level, comfortable with these systems; they empowered users of traditional, mostly print library collections and leveraged and reinforced the traditional philosophies and approaches to bibliographic control.

But these systems were not fundamentally revolutionary or transformational; they represent a process of modernization through automation, of measured evolution. The real revolution in access is just starting to arrive; this is going to be driven by the availability of massive amounts of content directly in digital form rather than print, and by the emergence of network-based computer systems that provide an environment not just for identifying content (which historically existed in print form and was used offline, independent of systems like online catalogs) but for its subsequent actual use and analysis within the access system. Indeed, the same computer systems that provide identification, access and an environment for reading and use may also serve as collaborative environments for new authoring . This is the new context for bibliographic control, and we ignore it at our peril; it will certainly reorder priorities for investment in bibliographic control practices and it will change the way that cataloging information, for example, is used and the purposes to which it is put in support of seeking relevant information.

I will focus here on methods of locating information that are distinct from, but complementary to, established practices of bibliographic description. A full understanding of these developments is essential in re-thinking bibliographic control in the new millennium, because they fundamentally change the roles and importance of bibliographic metadata in the information discovery processes.

There are three general approaches to identifying potentially relevant information:

- through bibliographic surrogates, that represent an intellectual analysis and description of aspects and attributes of a work; through computational, content-based techniques that compare queries to parts of the actual works themselves (or to computationally-derived surrogates for the works);
- through social processes that exploit the opinions and actions of communities that author, read, and evaluate works, and the information seeker's view of those communities of people involved.

The first approach is familiar, and forms the basis of catalogs and abstracting and indexing, and more recently online catalogs and similar systems. I will return to the question of how this changes in the new digital environment shortly.

The third approach is also familiar, in the form of book and article reviews, and suggestions from colleagues, and more recently citation indexing, but is now seeing a great creative expansion in the digital world, with its ability to create and aggregate world-wide communities of interest and to track the behavior of users within these communities. In this area we find fascinating and exciting current developments such as reccomender systems and collaborative filtering, which sometimes translate tracked behavior into implied ratings and which also permit the development of highly democratic, participatory and distributed explicit rating systems. We can also see here developments in trust and reputation management systems that begin to allow individuals to extend ideas about which opinions they trust and respect from limited and slowly changing circles of friends and colleagues to large dynamic global network-based communities that include many relative strangers. It is interesting to note that while this is a very powerful approach in support of individual information seekers, it is of much less use for intermediaries and for those concerned with the stewardship of collections.

The second approach is fundamentally new and indeed possibly only in the digital world, where techniques based on full text searching form the basis of todayÕs web search engines. The key point to recognize is that within a very few years virtually all new material, and an ever-growing amount of previously published material is going to be available in digital form as a routine matter. We need to recognize that in the new millennium, for digital materials, effective content-based computational techniques will be a very inexpensive, ubiquitous, default means of searching, available virtually the instant that the content is first distributed or published, and that powerful socially based approaches will also be widely available at little cost, as a byproduct of the authoring, dissemination and subsequent use of the works. The information identification support provided by human-based intellectual bibliographic control, which is intrinsically more costly and often available only after some delay following dissemination of a work, will have to compete with these other methods of finding relevant information, and do so with enough success to justify its costs.

It is worth noting some of the controversies surrounding full content searching and also worth recognizing some of its very real limitations.

Starting in the 1950s or thereabouts, a group of computational and information scientists began to develop a wide range of technologies to support effective full text retrieval without the use of bibliographic surrogates; their vision was that this would lead to far more effective and flexible searching and information location capabilities than bibliographic surrogates offered, and that ultimately it would also be far less expensive as the cost of computer cycles and storage continued to decline. Bluntly, if they achieved this vision, there would no longer be much need for bibliographic control, at least in support of information finding -- a very threatening prospect to many in the traditional library community. There were two major problems, however. Developing the technology to a reliable, robust level of maturity turned out to be extraordinarily difficult (as some people from the bibliographic control world enjoyed pointing out from time to time as the latest over-hyped technology developments surfaced). And even if the technology could be made to work, only an miniscule, insignificant proportion of the important literature existed in machine readable form so that the computation technology could be applied to it. Just about everything important was only available in print, while the researchers played with small, specially-constructed test databases.

Fifty years later, and after the investment of billions of dollars and countless years of human effort in research and development, the world has changed a great deal. The vision still hasn't been fully achieved (computers still have a lot of trouble deciding what texts are really "about", in a meaningful way, for example). But there is compelling evidence from full text searching systems (including web search engines) that content-based searching offers some capabilities that are completely unattainable through the use of bibliographic surrogates, and are often very valuable. Imagine being able to find every document that mentions a certain specific person, place or thing (right down to the passage in the document), to take one simple example. This is impossible with bibliographic surrogates (which weren't designed to solve this problem) but for many research needs it is absolutely revolutionary.

Researchers continue, appropriately, to push towards the vision and also to explore new ways that content-based retrieval can help information seekers; my personal view is that it will be a long time before they can replace human intellectual analysis by computation. But it is clear that current content-based systems complement traditional bibliographic control in supporting information seeking and provide capabilities that are not otherwise available. It is time -- indeed past time -- for the bibliographic control community to recognize the legitimacy of computational content-based retrieval and to understand its strengths and its contributions to information access, and also to look with an open mind at types of queries and classes of content where computational methods may compare favorably to bibliographic control based approaches, or may at least be "good enough", particularly given their very low cost.

As to the other objection, the paucity of content in digital form, as already discussed virtually all content is moving to digital form rapidly. The Web isn't' a test database -- it's a real-world collection of an enormous amount of information, some of it of great quality, importance and timeliness. There are some technical issues, and also some messy intellectual property issues (in part technical, in part legal and business) that will need to be resolved in order to make sure that the output of traditional publishing processes is available for indexing and searching by these computational systems (in the same way that it has been to catalogers, abstracters, indexers and reviewers), and this will take time and probably cause some considerable disruption and uproar along the way -- but this is another set of issues, for discussion another time. The key point is that we have now reached a "critical mass" of digital materials, and this will only grow, and this content will become available for computational indexing and retrieval.

There is one other essential point I must make here. Thusfar, while I have sometimes used the general term "content-based retrieval" what I have mainly been talking about is textual information. One of the great potentials of the digital environment is to elevate images, sound recordings, video, interactive simulations and other types of materials to a much more mainstream role in discourse, communication and the representation and capture of knowledge and of events than they have enjoyed up till now. We are already starting to see this happen; digital articles, term papers, or business communications can incorporate these nontextual components much more casually than their print predecessors. Tremendous amounts of audio and video are being routinely captured as a byproduct of various events and subsequently made available.

The best techniques that we have for making these kinds of non-textual materials available is to use human intellectual analysis to attach words to them (ideally within a structured descriptive or analytic context), and then to use these words as surrogates; much of this is essentially bibliographic analysis and control, or broader scholarly analysis, description and classification. Other techniques for using words to gain leverage on non-textual materials have a more mechanical character; transcribing talks, or creating closed caption tracks for video. There have been tremendous investments in technologies to make content accessible (mainly focused on the mechanical rather than intellectual processes), with varying results. Automated speech to text transcription has made significant strides in recent years, and continues to improve; this means that recorded speech, or the audio tracks of video materials containing recorded speech, can be automatically translated to text, and then methods developed for textual content can be applied (with some adjustments). Images and video have proven much more difficult -- in part because they can have meaning on so many different levels, and can concentrate a great diversity of meaning so intensely. Here intellectual analysis has been hideously labor-intensive and difficult; there are also fundamental conceptual problems about granularity and detail of description. I am reminded, for example, of the many ways and levels at which one can describe a painting of The Last Supper.

The most successful work on content-based image retrieval has, I think, occurred either in very constrained contexts (think about fingerprint matching, or face recognition) or has been limited to "vocabularies" very different from the way that most people think about images. (For example: I want images with lots of green on the bottom, blue on top, bits of yellow in the green -- this will retrieve meadows with flowers on sunny days, among other things, but it's not the way most of us usually ask for pictures of alpine meadows.)

For many kinds of nontextual materials, then, it seems that human intellectual intervention in the descriptive process is going to continue to be essential, at least for a considerable time to come. Bibliographic control of these materials is a part of this intellectual intervention to provide access. It's interesting to me that control of nontextual materials still seems to be one of the most complex and controversial areas, perhaps in part because there is a still not fully understood confusion of objectives in the work. But this will be a critical area as we think about the context for the new millennium; here the competitors to traditional approaches -- in particular content-based retrieval -- have more limited capabilities.

I've talked about three approaches to information access that, I believe, need to be viewed as complementary rather than competing, one of which is intellectual bibliographic control. The most effective ways to use the three approaches together is still a hard research problem (albeit one that forms an essential if uncertain context for any meaningful deliberations about the future of bibliographic control). But while this synthesis develops, it is also worth exploring possibilities for shared infrastructure among the three approaches, both as a way of encouraging synthesis (and indeed even dialog among the disparate communities that may help to advance such a synthesis) and as a means of leveraging investments. I offer three areas for exploration here which should be considered as another part of the context for the new bibliographic control.

First, we know that bibliographic control is not just about rules and practices. It also depends upon a rich and complex infrastructure of authority files and classification structures. Indeed, the other approaches also use infrastructure Ð for example, lexicons, dictionaries, gazetteers and similar tools for content-oriented computational techniques, and methods to manage identity, authenticity, and reputation in the case of socially-based systems. It will be important to determine how much of this infrastructure can be shared, and leveraged, among the three approaches, and what the practitioners of each approach can do to enhance this.

Second, we must recognize the democratizing and empowering character of the networked information environment; just as anyone can become a distributor of information with a global reach, anyone can become a describer of information. Quality and trust will be as much of a problem for description of content as it is for the content itself. Metadata itself is information, and we need to be able to decide when we choose to trust it; thus many of the same tools and techniques that have become relevant to the socially based discovery of information in the digital world will also become applicable in the production and use of bibliographic metadata Ð the linkage of metadata to identities through digital signatures, the management of identities through public key infrastructure, and the manipulation of reputation related to these identities. Thus we have a specific challenge in understanding how to connect and apply the infrastructure that is being driven by the social techniques Ð and indeed by much broader developments in the networked environment, such as electronic commerce Ð to bibliographic control.

Third, I believe that as part of the massive migration of content to digital form we are approaching a crucial point in standards-setting. Digital content isn't going to be simply text (or images, or sound); rather it is going to be complex structured objects that include both the "content" -- the text, images or whatever -- and also tagged metadata associated with the content. The particular metadata elements that are available will be important both for the automation of some traditional bibliographic control functions and for the support and enhancement of content-based and social information finding systems. All of the concerned retrieval communities need to have a voice in the discussions about standards in this area (along with other interested parties, such as those concerned with rights management, and the scholars who work with the materials). And I want to particularly highlight the linkage between these issues and issues about trust and quality -- for example, under what circumstances would bibliographic control practices countenance the automated extraction of metadata elements from a work into a bibliographic surrogate without human intellectual review and validation?

Clearly, there are opportunities for immediate and fruitful collaboration among the three communities of information finding practice, even as we strive to understand the deeper and longer-term questions about how to converge the contributions of the three communities, and how, in light of this convergence or synthesis, the practices of each individual community can be modernized, reshaped and made more effective.

We are entering a new world where content will be predominantly digital, and where it will be used, not just located, using electronic information systems. We cannot and must not attempt to map the future of bibliographic control without recognizing this. Continuing to ignore developments outside of the traditional scope of bibliographic control and to argue for business as usual -- and ever-growing funding to support business as usual -- runs the very real risk that our traditional practices may be discarded as unaffordable and of insufficient value in light of what the new technologies can offer. In my view this would be a tragedy; instead, we must concentrate on determining what bibliographic control practice can uniquely contribute, and where, when and how this contribution matters most. This means we must understand the changing context, and the economics, capabilities and limitations of the alternatives.

The economic pressures will be real as bibliographic control extends from print, where shared collaborative cataloging systems like OCLC have given us economies of scale in managing material that is acquired by many institutions, to special collections, where vast numbers of one-of-a-kind, unique items call for expensive original description. Worse, many of these items are non-textual, making them even more expensive to describe.

Finally, there is the problem of transition. Destiny may be digital, but we will be a long time reaching this destiny, and this long transitional period will call for careful management. We are already seeing print collections in our great libraries beginning to fade into invisibility for many patrons; materials available in digital form are so conveniently available, and so much more accessible through the range of retrieval systems when compared to print collections accessible only through bibliographic surrogates, and then further handicapped by document delivery considerations, that for these patrons the collection may as well only contain the digital content. While the amount of new material available in digital form is constantly growing, and there are major programs both in the noncommercial and commercial sectors to retrospectively convert print materials to digital form, this will be a slow process that will take many decades to complete. For these printed or other physical materials, bibliographic surrogates (and to some extent perhaps socially-based discovery systems) are the only means of access. What can be done to make them more visible, more accessible, to avoid partitioning knowledge into first-class (digital) information and second-class (physical) information? Bibliographic control carries a special, and heavy, burden here, and this raises serious questions about the allocation of resources for bibliographic control, and how to balance investments in bibliographic control and retrospective digitization.

This new context -- the emergence of cheap, ubiquitously available content-based retrieval approaches, and the great expansion of socially-based techniques for finding potentially relevant information -- leave us with a number of challenges in charting a future for the development of bibliographic control practices in the new millennium. What are the unique contributions of approaches based on human intellectual analysis? When is the use of intellectual analysis justified, and on what basis? What can we stop doing, or assign a lower priority to based on the assumption that content-based methods are available -- and how to our assumptions about the structure and format of the digital content that is available to these content-based retrieval systems (i.e. SGML or XML markup) shape our answers to this question?

Can we devise a spectrum of bibliographic approaches, with an accompanying spectrum of costs, to complement the content-based and socially-based approaches? Do we need to take the philosophically troublesome but perhaps pragmatic step of adopting different strategies for material that does or does not exist in digital form? How do we most effectively fuse the three approaches into information retrieval systems that are truly responsive to user needs?

The bibliographic control community cannot answer these questions alone. And they cannot shape their future without participating in a search for the answers to these questions. Redesigning bibliographic control for the new millennium will call for a new dialog among all parties and perspectives concerned with information finding that is grounded in a study of how the full array of tools and techniques now available can be applied to find information most effectively, and not in the inherent correctness or superiority of any one approach.

Library of Congress
January 30, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Martin Dillon

Former Executive Director of the OCLC Institute
Adjunct Faculty, OCLC Institute

Metadata for Web Resources: How Metadata Works on the Web

About the presenter:

From 1970 to 1985, Martin Dillon served on the faculty of the School of Information and Library Science at the University of North Carolina at Chapel Hill, where his research and teaching focused on topics in library automation and information retrieval. He came to OCLC as Visiting Distinguished Scholar in 1985. In 1986, he assumed the position of Director of the Office of Research, where he guided a staff of 30 in research supporting OCLC's mission of improving access to information. From June1993 until he became executive director of the OCLC Institute in January 1997, he served as director of OCLC's Library Resources Management Division, which is responsible for managing OCLC's Cataloging and Resource Sharing services.

As the inaugural director of the OCLC Institute, he led the Institute in forging new ways to facilitate the evolution of libraries through advanced educational opportunities.

Full text of paper is available

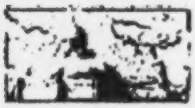
Summary:

This paper begins by discussing the various meanings of metadata both on

and off the Web, and the various uses to which metadata has been put. The body of the paper focuses on the Web and the roles that metadata has in that environment. More specifically, the primary concern here is for metadata used in resource discovery, broadly considered. Metadata for resource discovery is on an evolutionary path with bibliographic description as an immediate predecessor. Its chief exemplar is the Dublin Core and its origins, nature and current status will be briefly discussed. From this starting point, the paper then considers the uses of such metadata in the Web context, both currently and those that are planned for. The critical issues that need addressing are its weaknesses for achieving its purposes and alternatives. Finally, the role of libraries in creating systems for resource discovery is considered, from the perspective of the gains made to date with the Dublin Core, the difficulties of merging this effort with traditional bibliographic description (aka MARC and AACRII), and what can be done about the gap between the two.

Cataloging Directorate Home Page

Library of Congress Home Page



Library of Congress
June 27, 2000
Comments: lcweb@loc.gov

Metadata for Web Resources: How Metadata Works on the Web

by
Martin Dillon

Final version

1 Context of our Inquiry

First a brief, blunt statement of the context for our current activities. We are living through a revolution in knowledge representation. After a long and various evolution, knowledge representation settled into paper products for most of its output. Now we are shifting to digital forms for representing knowledge and to the Web as the primary distribution channel. This change will have profound consequences. There is little question, for example, that paper products will gradually be replaced by Web-accessible digital products. Is the Web here to stay? A premise of this paper is that the Web, or its evolutionary successor, will define the shape of our world for decades.

We are addressing questions concerning the cataloging function in this new world, a task that is complicated by uncertainties surrounding the future functioning of the library. Of necessity, one is closely tied to the other. Cataloging, after all, served libraries in a two-fold way: as a means of providing patron access to a collection of knowledge resources, and as a means of managing an inventory of such resources. Both of these were defined primarily as local functions applied to a local collection of paper products, which now will virtually disappear. How will this shift to digital knowledge change cataloging?

Addressing cataloging from the vantage offered above is a question that is central to the inquiry of the conference, but not of this paper. Even so, I want to make one point before I proceed here:

The library has to be reconceived as a unified cooperative, and cataloging has to be redefined as a function within that cooperative.

This fact seems painfully obvious but may still be worth stating, since the consequences that flow from it have never been worked out in any detail. Also, I regret to say, few of our colleagues have internalized this fact. Issues arising from managing Web resources from the collective viewpoint are not receiving the attention they deserve. Regrettably, most library activities directed toward providing Web access do so in isolation, acting to control an ocean tide with a teaspoon.

By contrast, cataloging in the paper world has benefited very much from the need to share work products among libraries, even though the results, from the individual library's perspective, were often viewed as primarily benefiting itself. One symptom of this local perspective, which can serve as an example of the split mentality of libraries, occurs when modifications are made to bibliographic records to conform to some local practice. These have long been a source of tension from the perspective of the global cooperative. I would argue, and have argued, that libraries would better serve their constituencies if they universally abandoned local variations in records in favor of record creation to serve a broader community.

In other words, where the bibliographic task in the paper world was defined primarily as the need to fit records into a local catalog, the new task we are designing our systems for is fitting surrogate descriptive records into a universal catalog for Web knowledge resources, with the added need, at least for the foreseeable future of having this catalog work congruently and seamlessly with the bibliography of the paper world.

That brings us to the task of this paper: how do we gain (bibliographic) control over knowledge resources on the Web? We have a new terminology to help us: resource description (or resource discovery) using metadata. I will address the reasons why I distinguish discovery from description below, when we get to the Dublin Core, but first I want to discuss the concept of metadata.

2 Definitions of Metadata

Metadata is a recent coinage though not a recent concept. In today's jargon, metadata is data about data: information that communicates the meaning of other information. As nearly as I can tell, the term has come to prominence in our context only with the Web, dating from the early 90's, where it surfaced in the face of a newly recognized need: resource discovery on the Web. (See below in the Notes section, METADATA, the trademark)

We find the first oblique reference to metadata in the "HyperText Markup Language Specification Version 2.0," which discusses "meta-information" in the header section of a HTML document:

Meta-information has two main functions:

- *to provide a means to discover that the data set exists and how it might be obtained or accessed; and*
- *to document the content, quality, and features of a data set, indicating its fitness for use.*

(/http://www.w3.org/MarkUp/html-spec/html-spec_toc.html)

The first of these bullets targets resource discovery; the second targets resource description. The first mention I can locate for the term "metadata" used in this sense occurs in the Geospatial community and its efforts to define resource description systems for geospatial data: "Content Standards for Digital Geospatial Metadata Federal Geographic Data Committee," dated June 8, 1994.

At the risk of adding to the confusion surrounding this term, I would like to expand the concept of metadata to include a second type: data labeling. Indeed, this type of metadata can be viewed as primary, as more basic than resource description. I would like to elaborate briefly both forms of metadata.

Metadata as tags

The most common form of this type of metadata arises from the use of tags to characterize the content of fields. This kind of metadata has a great variety of uses. It is found in all information forms: survey instruments, purchase forms of all sorts, and yes, tax forms. What all of these forms have in common is that they contain labeled fields: a text definition followed by a blank space. The different fields are meant to be filled in and later processed. Labeled fields of this sort are also found in all commercial record keeping, most particularly in the world of electronic data processing, where such standards as EDI have been promulgated to allow information exchange among cooperating commercial firms.

Our focus is exclusively on fields defined by the tagging that occurs in markup languages. SGML was the first of a series of standards that were initiated in the late 80's and has recently culminated in XML. The tags in these systems occur in pairs; each pair defines and delimits a field, with the contents of the field occurring between the two tags. All markup languages (SGML, HTML, XML) make use of this kind of metadata. A simple example:

```
<title> Any title </title>
<publisher> Amazon.com </publisher>
<price> $12.50 </price>
```

Each field (or element in the terminology of markup languages) has a start-tag (<...>) and an end-tag (</...>). The character string within the brackets identifies the field; the area between the start-tag and the end-tag contains a character string that is the value of the field. In the above example, the pairs of bracketed names: <title>, </title>; <publisher> ,</publisher>; and <price>, </price> are the metadata; these metadata convey information about the character strings within each of the pairs. The data thus described are 'Any title', 'Amazon.com', '$12.50'.

This kind of metadata has the advantages of simplicity, machine and human readability, and great expressive power, as HTML has demonstrated in the Web environment. Until recently, HTML tagging has been used to "mark up" all Web content, promiscuously conveying information about formatting, linkages and descriptors.

Metadata as descriptors

But here's the kicker: In our example above, the strings occurring between each start-tag and end-tag are also data about data: they are also metadata. In the example, they are about a publication and are therefore bibliographic in nature.

When discussed in a Web context, the term "metadata" can refer to either type: the tagging system that defines a set of fields and its contents, or the contents of certain fields that act as descriptors for other resources. This duality can create confusion and it doesn't help that the same string of characters can act as metadata on one level, and data on another, depending on the perspective being used.

2.2 Metadata on the Web

In tackling the problem of providing descriptive surrogates for library-related Web resources, we have to be concerned about both kinds of metadata for the following reason: the tagging systems for Web pages, and the conventions and standards for processing them, create the context within which library practices reside; the infrastructure of the Web is driven by them and creates the opportunity for us to build within it a means to achieve our own ends. Since it is the crucial underpinning for our own efforts, before we focus on resource description, we need to discuss briefly the general use of metadata tagging in the Web environment. Such tagging has had a wide variety of applications on the Web independent of libraries. Each application has had its metadata standard proposed, debated, implemented and sometimes abandoned. We will consider some as preparation for our library applications.

General Metadata Systems

By general metadata system, I mean a methodology for fully characterizing all of the data for an application. The two primary examples of such general systems are:

- *"The Meta Content Framework Using XML," a proposal submitted to the World Wide Web Consortium (W3C) in June 1997, Netscape's major contribution to the metadata initiative.*
- *The "Channel Definition Format," submitted in March 1997, is Microsoft's major contribution to the metadata initiative. It "extends XML and Web Collection work that the W3C" has worked on. CDF is the "industry's first" channel framework for push technology on the Web.*

It will not benefit us here to do more than mention general metadata systems other than to state that their primary aim is to enable the precise mark up of data streams for system interoperability.

Resource description

Problems of resource description have pervaded the Web since its beginnings. Not surprisingly, however, metadata for resource description have not always been provided explicitly in Web pages. The "Head" section of the HTML Standard was introduced in version 2.0 (early 1994) when the Web was 2 years old. It included the "Meta" element for the first time with such attributes as "title". Metadata in this form proved very popular, with its use growing very rapidly. By 1998, 70 % of public Web sites made use of them, with an average of 2.75 meta fields for each site that used them. ("Web Characterization Project: An Analysis of Metadata Usage on the Web," Edward T. O'Neill, et al) (www.oclc.org/oclc/research/publications/review98/oneill_etal/metadata.htm)

This form of resource description, our primary topic here, engages virtually all Web users, and ranges from search engines and directories of all types to the identification and discovery of special interest communities.

PICS and other content controllers

The Platform for Internet Content Selection (PICS), an activity related to resource description, both historically and practically, is based on the desire to filter or restrict access to materials of certain types. The most obvious is pornography and the filter or restriction is with respect to juvenile access; but there are many cultures that wish to restrict access to other materials, mostly of a political nature. How to do this within a Web context is the primary question, and the answer is through characterizing the content of resources from this vantage. The O'Neill study noted above does not find much use of PICS tagging. See (www.w3.org/TR/REC-DSig-label/#DSig_1_0_Overview) or (www.w3.org/PICS/) for further information on PICS.

Commerce - BizTalk and SOAP

From a Microsoft June, 1999 press release, "the BizTalk Framework is an open specification for XML-based data routing and exchange. The BizTalk Framework makes it easy to exchange information between software applications and conduct business with trading partners and customers over the Internet." SOAP, the "Simple Object Access Protocol" developed by Microsoft, "is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined datatypes, and a convention for representing remote procedure calls and responses." (Taken from the document submitted to the W3C recommending the formation of a working group for Web protocols (Simple Object Access Protocol (SOAP) 1.1, W3C Note 08 May 2000) See (www.microsoft.com/biztalk/) for details.)

Depending on whom you talk to, BizTalk and Soap are either an alternative to the Resource Description Framework (discussed in the next section) or a complement to it. In either case, the existence of both, with neither giving any evidence they are aware of the other, is indicative of the diffuse effort that reigns in the Web arena over how to solve the need for interoperability and data exchange among distributed applications that are the norm on the Web.

Rights Management

And one such distributed application is the management of intellectual property rights on the Web. The need is to protect intellectual property rights on the Web and enable commercial publishers to control effectively the electronic transfer of such rights. The International DOI Foundation, in collaboration with commercial publishers, is responsible for advancing the definition and uses of the Digital Object

Identifier (DOI(r)), and is among the leaders in the endeavor to manage property rights. The DOI is "an identification system for intellectual property in the digital environment." Its principle objective is "to develop automated means of processing routine transactions such as document retrieval, clearinghouse payments, and licensing." (http://www.doi.org/index.html) Metadata arises in this context as a means to identify, describe, and allow the tracking of all manner of intellectual property on the Web, to protect it from misuse, and to enable its creators to be properly remunerated.

Although part of the objective of DOI Foundation is to provide a basic resource description to accompany the DOI identifier, much like the elements of the Dublin Core provides, it is noteworthy that no mention of the Dublin Core occurs on their site.

2.3 RDF: the Resource Description Framework

Before concluding this section on general issues dealing with metadata on the Web, and before turning to the metadata of resource description, I would like to discuss briefly the relevance of the Resource Description Framework, henceforward referred to as RDF. The best overview of what RDF is and what it is to be used for remains Eric Miller's "An Introduction to the Resource Description Framework" appearing in D-Lib Magazine (http://www.dlib.org/dlib/may98/miller/05miller.html). From the abstract, "The Resource Description Framework (RDF) is an infrastructure that enables the encoding, exchange and reuse of structured metadata."

From the W3C RDF FAQ:

> *RDF emphasizes facilities to enable automated processing of Web resources. RDF metadata can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities; in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; in content rating; in describing collections of pages that represent a single logical "document"; for describing intellectual property rights of Web pages, and in many others. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.*
> *(http://www.w3.org/RDF/FAQ)*

It is not clear as yet what relevance RDF has to the library world; more broadly, and perhaps causative, it is not clear as yet what relevance RDF will have in the Web. The attitude of Web practitioners toward RDF varies greatly. At one end of this spectrum is the W3C community, which maintains that RDF will provide the mechanisms to solve many of the interoperability problems in the Web. At the other end is Microsoft, which, so far at least, has exhibited a deafening indifference to RDF. The latter attitude is manifested by a total avoidance of its use within Microsoft's product line, and is an almost reflexive corporate reaction to any standard not created by Microsoft itself. If the Microsoft reaction is indicative of the low rate of adoption generally, then RDF is in trouble.

What does the success or failure of RDF matter to the library community? From the perspective of the library world acting within the boundaries of its own community, successful resource description standards and methods are possible without an RDF. Moreover, as with many other Web developments, RDF will succeed or fail based on the practices of the larger world outside libraries. As is so often the case with emerging standards, watchful waiting is probably the best approach.

The Future of XML

The future of RDF is tied closely to the emergence of XML. What is the future of XML? First and foremost, it appears clear as of this writing that HTML as the markup language of choice for the Web will eventually give way to XML. XHTML, a recent variant of HTML, was designed to provide a bridge between the two. I have heard numerous optimistic predictions about the pace of this evolution, all of them wrong so far: installed systems are always slower to give way than one would wish. Two milestones will be worth watching for: when half of all new Web pages being written are in XML; and second, when half of all the pages on the Web are in XML. Neither will occur any time soon, certainly not in one year, very probably not in two.

Below, I discuss the impact of this change on library issues. The primary issue, however, remains that we are at the mercy of the general Web community in these areas. Progress will occur at a pace dictated by the needs of large movers on the Web, influenced to some degree by the general problem of resource discovery experienced by all Web users, and also by all of those other applications awaiting an effective solution. If this brief consideration of metadata uses on the Web accomplishes anything, I hope it communicates the diversity of communities engaged in providing standards and also the lack of cohesive efforts and results that have been achieved thus far.

3 Metadata Standards for Resource Description

Now that we have gotten through the preliminaries, we can turn to our major topic: metadata used for resource description on the Web. It may help clarify Web efforts to touch first on standards that fall under the general topic of resource discovery but were not designed specifically for Web resources. They include such standards as those developed by the Consortium for the Computer Interchange of Museum Information (CIMI), those standards whose development is funded or directed by the Federal Geographic Data Committee, mentioned above in relation to the term "metadata"; and the Government Information Locator Service (GILS), now used to provide access to government documents. These three standards were developed outside the library community. Examples of metadata standards developed within the library community would include the Text Encoding Initiative (TEI) and the Encoded Archival Description (EAD), which were created using SGML and pre-date the Web, but which have since been converted to XML for use within the Web. Links to all of these are provided in the "Resource Section" below. *None of these can be said to have arisen because of the Web, nor was their initial focus on Web resources.* Rather, they use metadata to provide finding tools for patrons in their respective applications. They are more or less parallel to systems of MARC bibliographic records: they are systems constructed to provide descriptions for various classes of objects in the areas of application, ranging from the contents of museums to archived papers. As with almost everything in today's world, the Web is increasingly important as a mechanism for meeting the needs of users by connecting them to resources, whether those resources are available for use on the Web, or only described through the Web and require further action in the non-web world. Items purchasable through the Web fall into the latter category.

The major Web mechanism for connecting user to resource is the search or directory service. Both make use of resource descriptions either to allow the user to perform a search or allow browsing. Typical and relevant is an OPAC search to locate a book, or a similar search on Amazon.com. In neither case is the book itself available on the Web, at least not yet.

To the extent that the standards referred to above deal with objects not directly usable through the Web, they fall outside my concern here because I would like to focus exclusively on Web resources.

One final point: This distinction between Web resource and objects outside the Web may appear somewhat arbitrary. While deploying metadata systems, there is often an overlap between the two. CIMI, for example, has been and is a very active participant in the Dublin Core community, which is responsible for creating the Dublin Core, the preeminent resource description standard in the Web environment. CIMI participates in the Dublin Core at least in part because so much of its resource description activity is manifested in some form on the Web. Increasingly, it is possible to link to images of museum objects on the Web; these images are Web resources par excellence, and thus very much a target of the Dublin Core community. The same can be said for archival information covered by the EAD community: one day all of these materials may be accessible on the Web.

The needs of these various communities for resource description capabilities create a challenge for standards bodies seeking to create tools that can accommodate them. In their complex combinations, they raise questions about the nature of surrogate records. The Web is so universal, so all-encompassing, that we look toward a time when everything will require its Web surrogate to find its user. This aim implies a need for surrogate languages with great expressivity. The ambition of standards such as XML, RDF and the Dublin Core is to achieve this level of expressivity.

We can now turn to the Dublin Core and assess its attempt to accomplish the lofty aims set forth here. And we will encounter a regrettable limitation on the human condition: when we try for too much, we often deliver too little.

3.1 The Dublin Core Metadata Standard

The standard central to our purposes is the Dublin Core, which arose within the diverse standards creation activities of the mid-90's. From the outset the Dublin Core had as its focus resource discovery on the Web. As stated in a 1998 IETF document, "The Dublin Core Metadata Workshop Series began in 1995 with an invitational workshop which brought together librarians, digital library researchers, content experts, and text-markup experts to promote better discovery standards for electronic resources."

([RFC2413] Dublin Core Metadata for Resource Discovery. Internet RFC 2413. (http://www.ietf.org/rfc/rfc2413.txt))

"Discovery standards for electronic resources" - as noted earlier, I have used the phrase "resource description" instead of "resource discovery" because description is more general, and in my view more accurately characterizes what is required. One may claim that an effort is restricted to resource description, but if one does not deal with user needs effectively, no justification will satisfy. Resource discovery is impossible without resource description; adequate resource description assures effective discovery. The difference is as basic as the difference between a keyword search and an adequate display of results. The former allows discovery; the latter, based on resource description, allows effective selection from an extended list. I will elaborate this more fully below when we discuss alternatives to cataloging.

In library terms, the Dublin Core is a simple system for cataloging Web resources, no more, no less. And it should be judged from that perspective.

3.1.1 Issues with the Dublin Core

Many issues surround the primary question of the effectiveness of the Dublin Core, and I would like to list and discuss them briefly.

Degree of completeness

Unfinished - the most serious problem of the Dublin Core to date. The first official version of Simple Dublin Core was available in 1997 after 2 years of discussion and debate. The first published version of a qualified Dublin Core was made available in July of this year. It is obviously incomplete, with no qualifiers being offered for the Creator, Contributor, Publisher elements. As yet no one has been able to provide documentation, extensibility rules or implementation guidelines for a qualified Dublin Core. What this has caused in the intervening years is the development of various community versions of qualified Dublin Core's. What this has also caused in the intervening years in every community attempting to apply the Dublin Core to a collection is endless debate over what the various elements mean and how they are to be used. What this has also caused in the intervening years is very slow adoption of the Dublin Core as a standard for resource description for the Web. (Again, see O'Neill's report cited above for statistics.)

Institutional support

Lack of institutional support is not surprising given the degree of incompleteness of the Dublin Core. CORC (Cooperative Online Resource Catalog), a new service from OCLC introduced in July of this year, which incorporates the newly published qualified Dublin Core, is a strong step in the right direction, but much more is needed, including a standards body and procedures for evolving and changing the Dublin Core.

Documentation

Documentation, of course, must follow on a published standard and can't precede it. After the recent release of a qualified Dublin Core, it may be possible now to provide at least some usable documentation.

Implementation guidelines

As yet there is no direction on how to implement the qualified Dublin Core in HTML or XML, though this may change at any time.

Extensibility rules

There is as yet no precise direction on what counts as an allowable extension to simple Dublin Core, or what syntax extensions must conform to. The absence of a clear definition of the syntax of qualifiers continues to make implementation guidelines difficult if not impossible to achieve. Sufficient for this purpose may be the Dublin Core Metadata Initiative (DCMI) publication prepared by the DCMI Usage Committee, which "describes the principles governing Dublin Core qualifiers, the two categories of qualifiers, and lists instances of qualifiers approved by the Dublin Core Usage Committee." ("Dublin Core Qualifiers," July 2000) (http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm)

In that document, two kinds of modifier for elements are recognized: Element refinement and Encoding scheme. The first is characterized by such modifiers as "created" for the date element; the second by "LCSH" for the Subject element, and "URI" for the Identifier element. For explanations and further examples, please refer to the official publication cited above where all qualifiers defined for the current version are presented in a table. I have gone into this level of detail here concerning acceptable qualifiers for Dublin Core because I explore a problem with respect to them in the next section.

3.2 Other Issues

The above list of Dublin Core issues may be transitory. Indeed, it is possible that some of them will be removed or at least alleviated by the DCMI July, 2000 publication cited above. What if they were all fixed? Would our need for a resource discovery standard for the Web be satisfied? There are two general areas of concern that I can see. First, if we generously assume that the Dublin Core in its current form is approximately finished, and that its major focus is on "document-like objects", how close is it to an acceptable standard? Will tweaking over time and through experience in its use gradually provide us with a standard we can live with? Or are there major fissures that must be bridged? Second, does the architecture of the Web require a standard that goes beyond an object-attribute model for resource discovery? I would like to discuss each of these briefly.

3.2.1 Difficulties with the current form of the Dublin Core

The current structure of the Dublin Core limits its usefulness in critical ways. As outlined above, Qualified Dublin Core currently allows an element to have two modifiers: the first is considered to be a refinement of the element; and the second, the encoding scheme, is considered to modify the value of an element. The distinction between these two types of modifiers, and others that might be used, have been the source of much discourse within the Dublin Core community, one cause of the delay in completing a draft of a qualified Dublin Core. My problem is fundamental and practical and can be expressed by citing, as examples, what I consider to be serious weaknesses in two Dublin Core elements: the Creator and the Relation elements.

Creator element (and Contributor and Publisher as well)

What is needed for a Creator element (or what I would like to see it have!) is a structure that provides for the name of the creator as its value, a modifier that states whether the name is corporate, personal, or geographic, and a further modifier that is a URI pointing to an authority record for the name. (All modifiers, like all elements in the Dublin Core, are optional.) The capability of attaching a URI to a Creator element would not only obviate the need to include supplemental Creator information such as an email address (which many have recommended, and which I consider to be highly undesirable), but it would also allow, and thus encourage, a far more effective means of authority control in the Web environment. The fundamental Web mechanism is the link; a Creator field should link directly to the authority record . What could be more natural, desirable, powerful? My understanding is that a group is investigating how to handle authority linkages with the Dublin Core; I hope this solution is still a possibility.

Relation element

The Relation element poses a similar problem arising from the same structural cause: more modifiers are required to give the Relation element what it needs for effective use. A relation element contains information about a "related item". Three pieces of information are required for this element to be a useful Web construct: the name of the relation ("Is part of", "Is version of", etc.), the name of the item (in the simplest case, a title), and, when available, a URI to get to the item.

Under the current structure, we can provide either a name or a URI, but not both.

There is a solution to both of these problems and one in accord with the essence of the Web: define as part of any Dublin Core element a pointer element for "additional information."

3.2.2 Difficulties with the Object-attribute model

Web Resources: the medium is the message

Marshall McLuhan's famous dictum, "the medium is the message", recommends caution in how we understand the workings of a new medium. Our new medium is the Web and what McLuhan meant, I suppose, and what has application here, is that the characteristics of the medium often have greater impact or influence than the actual content. We are moving from a print culture to an online culture. In the present context, the characteristics that are most at issue involve the change from "collections" and "objects" to ... pages and pointers? Resources? To what? And why do we care?

We care for a number of important reasons. It can be argued that AACR2 cataloging is, by its very nature, tied to physical objects, and when we move into a world without physical objects, the target of the cataloging effort becomes fuzzy or without boundaries. This lack of definition may create insurmountable obstacles to the effective application of cataloging principles and practice. I subscribe to this view without understanding it fully, and I will attempt in what follows to explain why.

Objects vs. resources vs. whatever

Back in 1992 when we undertook to examine "access to Internet resources", (a project reported in "Assessing Information on the Internet: Toward Providing Library Services for Computer-Mediated Communication," (Spring 1993), Internet Research, 3(1) 54-69) we played a simple trick on ourselves to sidestep the issue I want to discuss here. The trick was tactical and was necessary at the time for us to make progress: *we restricted our investigation to "document-like objects" on the Internet*. We chose this route to make progress because our first meetings had become bogged down in discussions about what sorts of things were on the Internet, how they differed from documents, and what the implications for cataloging were. After a few rounds of profitless discussion and no progress, by fiat we restricted our focus.

What is the essence of the problem? I believe it is in the notion of object-hood and how that notion does not translate very well to the Web. Consider first one of the basic principles of Anglo-American cataloging: the item in hand. Much depends on this concept including a well-defined boundary for the cataloger in the cataloging process. Of course, even in our workaday world where the cataloging target is a discrete physical package, there are severe problems that must be overcome. Many of these arise because of the differences between the class of objects related to what is referred to as the work and the classes of objects in the work's various manifestations. Questions concerning differences between one class of manifestations and another are legitimate and deserve the attention they receive; how they are resolved determines, among other things, when a new record is required for an item in hand, and when an existing record will suffice. Though important, discussions of these issues have often been unsatisfying. It may be that the problems they pose are fundamentally intractable, that cataloging offers a means for creating round holes into which through various compromises we force a collection of square pegs.

In the world of physical objects, part of the problem certainly is the oversimplification encouraged by the illusion that the ground is solid beneath our objects. One example, long a favorite with me, has to suffice. A trivial pursuit question:

Category: cataloging. What is the smallest difference between two books that will lead to the creation of two different bibliographic records?

In more general terms, how big does a difference have to be between two objects to justify the creation of a second bibliographic record? We are touching on the question for which the Dublin Core "1:1" rule offers the answer. And the answer may be unwise, wrong-headed or otherwise misguided, but it assumes object-hood: one object generates one record.

The problem I want to address is the following: is object-hood an effective metaphor for successful resource description in the Web? Please remember that we are not dealing with absolutes, either all or none. In the print world, object-hood has its limitations: the concept of serial was invented to deal with one of them and the discussion above exposed a more subtle definitional problem in dealing with monographs. On a scale of 1 to 10, we could say that for monographs, item-in-hand object-hood is 9.8 successful. What degree of success are we likely to achieve using object-hood as the basis of cataloging on the Web?

The "1:1" rule assumes objects as a given. Its primary purpose is to deal with problems arising when more than one manifestation of the same work exists. Simple examples will suffice: differences in format, say PDF and RTF; or different representations of some object, say image or Html. This oversimplifies but does no harm here, because the very notion of recognizable objects is undermined in the Web.

From the perspective of managing those Web resources that are of interest to the library community, the question becomes: how many conform comfortably to the notion of an object; conversely, how often will an assumed object-hood get us into trouble? Is the use of an object as the underlying metaphor a useful fiction? Or is it more apt to get us into a heap of trouble?

It is always useful to bring forward examples from the print world when they are available to shed light on difficulties like the current one. Two occur to me. The first is the practice of faculty creating a collection of readings gathered from disparate sources as a quasi text book for a course. I have never heard of anyone advocating that libraries catalog such an object. But why not? Surely, surrogates for such objects would be useful if the table of contents were included. Would not others teaching similar courses benefit from having access to the description of the book?

Perhaps a more apt example, certainly a more recent one, is the possibility of anyone creating his or her own book by gathering pieces and parts from a large database of books, whose contents are themselves stored and accessible in parts. Not only chapters and sections could be extracted, but pictures and tables and any other pieces at the whim of the purchaser. As depicted by Lisa Guernsey, "Under this model, books have not only turned into streams of electronic bits that are downloaded to hand-held devices or printed on demand. They have also turned into databases -- pools of digital information that people can extract and combine on their own terms." (From "Books by the Chapter or Verse Arrive on the Internet This Fall," NY Times, July 18, 2000)

Clearly, the results of this process are outside the scope of the cataloger.

I would argue that a Web resource is often much more like a fluid, multi-dimensional, multi-layered, constantly changing complex of things and relationships than it is like a simple object. Web resources do not have tidy boundaries.

Web Resources

It is necessary to probe this issue further. Web resources are different from monographic objects in ways that profoundly change the cataloging problem; this difference is growing: more of the Web can be thus characterized and the distance between such resources and the monographic object is growing.

Most simply, the problematic characteristic of the Web resource is one of extent: it is difficult, if not impossible, to define the extent of a Web resource, to state where it begins and where it leaves off. Try defining these terms: Web page or Web site. They are used ambiguously on the Web and in the literature. Moreover, what relation do they have to the terms: file, directory, or server? The vagueness of the terminology in this area is symptomatic of the vagueness, in physical terms as well as conceptual terms, of the underlying concepts.

Before we can catalog something, we have to know what we are talking about.

4 The Role of Libraries in Web Resource Description

We also have to know what we want to accomplish. Barbara Baruth, in a recent article in American Libraries ("Is Your Catalog Big Enough to Handle the Web," August, 2000, pp. 56-60) explores the question of the library's role in resource discovery on the Web. She asks, "Will the impressive second-generation search engines out now or third-generation engines now incubating make the idea of quality-based services such as CORC obsolete?" Future search engines, she continues, may be able to do a fine job, "scouring the net and bringing back tailored results." And finally she asks the sixty-four dollar question, "Is it possible that manual efforts to explore, evaluate, and catalog the vast reaches of the Internet just can't compete [with these advanced search engines]?"

What is the library responsibility with respect to providing access to Web resources? What is its role, and how should it carry out this role? Until we provide credible answers to these questions, it is not possible to chart the future course of libraries, and secondarily, cataloging. Even if we agree with Barbara Baruth's assessment that search technology will improve sufficiently to eliminate the need of human resource description, how long will this take? I am always suspicious, and I recommend this scepticism to all, when delivery is promised of technologies that are not yet in beta test. Experience tells us that the promised date almost invariably stretches into the future.

Let me state my own view: I see no hope that searching alone will replace the need for human cataloging in the forseeable future, that is, the next 5-10 years. Here are some reasons for my view:

Wrong, obscure or missing information

Searching is similar to automated cataloging in that neither can overcome the absence of data inferable from a resource, and Web resources will not evolve stable self-describing mechanisms for a long time, if ever; such mechanisms are not yet even being broadly discussed. Desired characteristics such as creation date, revision date, and expiration date, just are not easily available from most Web resources. Inappropriate titling, weak or absent content descriptors - we can go on and on. The absence of these descriptors, or their presence in corrupt or unrecognizable form, within a Web resource corrupts the results of any searching; and we can expect such problems to grow for a long time rather than abate.

Authority control

The problem of coordinating and differentiating names, a modest source of difficulty within the controlled environments of the library catalog and the commercial publishing world, becomes a nightmare on the Web. All of the usual suspects are involved: personal names, corporate names, geographic names, subject descriptors; all now compounded by language and character set confusion on an immense scale.

Selection

Finally there is the issue of selection. The Web now has over a billion pages, whatever that means. The task of culling from this huge morass the population of stuff that we want to search is almost overwhelming. It can only be accomplished by an equally huge, ongoing effort of thousands of people, effectively coordinated by well-designed online systems.

5 Conclusions and Recommendations

Let me take a final quote from Barbara Baruth's article cited above: "The future of library systems architecture rests in the development of umbrella software that digests search results from rapid, coordinated searches of a variety of disparate databases." That is, the job of resource discovery will be accomplished primarily through software directly acting on Web resources without benefit of human intervention, particularly of the cataloging sort. I disagree with this position on a number of grounds, not least that I believe that searching alone will reach a point of diminishing return (may have already). A second, library-centric reason is based on the assertion that if the library role can be encapsulated by such search engines, we can dispense with libraries forthwith: this functionality can be provided by software firms and distributed directly to patrons either as clients or by glitzy Web portals.

I would argue that it is the responsibility of the library to provide effective access to knowledge resources on the Web. If the various commercial services can adequately accomplish this library goal, let's get on with other worthwhile knowledge management tasks required by our patrons. Barbara Baruth is certainly not alone in the belief that such services are rapidly succeeding in this goal. A parallel here is the dependence of libraries on abstracting and indexing services, which provide tools for accessing the journal literature. Nothern Light and Google are Web versions of the same idea.

Let us assume that library intervention is required for successful access to Web resources of interest to patrons. For those resources that are roughly equivalent to documents in the physical world - self-contained, more or less static - the cataloging task emerges in much like its historic form. No small task because there are a great many such objects. Let us continue to ignore that other class of resources, those whose object-hood is in question.

How should libraries provide access to document-like knowledge resources on the Web? If the library community decides that it is necessary to establish a form of bibliographic control for such objects, three paths are open:

1. Use or adapt MARC/AACR2
2. Start fresh creating a library metadata system with the same aims as the Dublin Core
3. Use or adapt the Dublin Core

I will discuss each of these briefly.

Use or Adapt MARC/AACR2

There may have been a time when this was a useful direction to take but it is long past. The result of such an exercise would have many of the attractive attributes of the Dublin Core, particularly its simplicity and flexibility.

Start Fresh

A fresh start, guided by the lessons learned from the long parturition of the Dublin Core is an intriguing idea. But is it realistic? Can the library profession manage the rapid creation and deployment of such a standard? Nothing in our history encourages optimism.

Use or Adapt the Dublin Core

We are left with this final option. It is more likely that we can make progress by either using whatever version of the Dublin Core is current, or, far better in my view, attack the problem of creating a library-specific variant of the Dublin Core that suits the aims of the library. The criticisms of the Dublin Core offered above provide at least a starting point for what such a variant might look like.

As a final point, I would only strongly recommend that at least one action be taken fothwith: that a MARC version of the Dublin Core be developed, with appropriate instructions and examples. The work products of such a MARC include at least the following:

- The list of fields and sub-fields defining the MARC Dublin Core record, including an indicator that the record is a Dublin Core record.
- Necessary documentation with appropriate examples.
- A definition for a MARC input screen to guide local system vendors and utilities.
- A plan to urge cataloging utilities to incorporate this style of record into their editors.

I am not suggesting a multi-year project; my guess is that this work effort could be accomplished satisfactorily in a matter of a very few months.

This MARC version and its accompanying documentation would be suitable for use in library OPACs, if desired, and would be directly convertible to and from any database of Dublin Core records. The advantages of doing this are obvious. It would immediately communicate to thousands of catalogers the essential nature of the Dublin Core and equip them to make use of existing systems and software to create resource descriptions for Web resources. Would this be a solution to our problems? No, but it would put us in the game as it is defined in today's Web world. Consider where we would be today if a library-defined version of the Dublin Core existed 3 years ago. If the MARC Dublin Core was adopted and vigorously applied by thousands of libraries, we would be far better positioned to serve the Web needs of library patrons and Web knowledge access would be far different and far better.

6 Notes and Sources

6.1 METADATA, the trademark

Thanks to Rick Pearsall, FGDC Metadata Coordinator, I learned that the term "Metadata" was trademarked in 1986 by The Metadata Company (The Metadata Company, http://www.metadata.com). Its invention is credited to Jack E. Myers who is said to have coined the term in early summer of 1969. The trademark should be written with capital letters and should be distinguished from both "meta data" and "meta-data".

6.2 Metadata System Examples

6.2.1 Content Standard for Digital Geospatial Metadata (CSDGM)

http://www.fgdc.gov/metadata/contstan.html

An outstanding example of metadata definition is that developed for Geospatial data and mandated by the Federal Government.

The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data, to determine the fitness the set of geospatial data for an intended use, to determine the means of accessing the set of geospatial data, and to successfully transfer the set of geospatial data. As such, the standard establishes the names of data *elements and compound elements to be used for these purposes, the definitions of these data elements and compound elements, and information about the values that are to be provided for the data elements.*

As stated in the documentation for the standard, "The first impression of the CSDGM is its apparent complexity; in printed form it is about 75 pages long. This is necessary to convey the definitions of the 334 different **metadata** elements and their production rules. Do not let the length dismay you." (http://www.lic.wisc.edu/metadata/metaprim.htm, 'Metadata Primer -- A "How To" Guide on Metadata Implementation') If you are dismayed by its length and complexity, join the crowd!

6.2.2 U.S. Geological Survey. Government Information Locator Service.

URL: http://www.gils.net/

A useful source document is available through the U.S. National Archives and Records Administration (NARA). Guidelines for the Preparation of GILS Core Entries.

URL: http://www.ifla.org/documentslibraries/cataloging/metadata/naragils.txt

6.2.3 The Consortium for Interchange of Museum Information (CIMI)

From the introduction at the site: CIMI (the Consortium for the Computer Interchange of Museum Information) is committed to bringing museum information to the largest possible audience. We are a group of institutions and organizations that encourages an open standards-based approach to the management and delivery of digital museum information.

http://www.cimi.org/

A useful overview is provided in, "The use of XML as a transfer syntax for museum records during the CIMI Dublin Core test bed : some practical experiences."

http://www.cimi.org/documents/XML_for_DC_testbed_rev.doc

6.3 Other Sources

6.3.1 INDECS: interoperability of data in e-commerce systems

An international initiative of rights owners creating metadata standards for e-commerce - "putting metadata to rights" . INDECS provided the metadata model for the DOI. The site has links to background information on the INDECS project and its results.

http://www.indecs.org/index.htm

6.3.2 Digital Library: Metadata Resources -

The single best source for all aspects of resource discovery metadata

http://www.ifla.org/II/metadata.htm

6.3.3 The Resource Description Framework

Dave Beckett's Resource Description Framework (RDF) Resource Guide

http://www.ilrt.bris.ac.uk/discovery/rdf/resources/

The offical source document for RDF defines it as

Resource Description Framework (RDF) is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasizes facilities to enable automated processing of Web resources. RDF can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities, in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library, by intelligent software agents to facilitate knowledge sharing and exchange, in content rating, in describing collections of pages that represent a single logical "document", for describing intellectual property rights of Web pages, and for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.

http://www.w3.org/TR/PR-rdf-syntax/

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Sarah E. Thomas

University Librarian
201 Olin Library
Cornell University
Ithaca, NY 14853-5301

The Catalog as Portal to the Internet

About the presenter: Sarah Thomas came to Cornell University in August 1996 as the Carl A. Kroch University Librarian. In a career spanning over 25 years, Thomas has cataloged books in Harvard University's Widener Library, taught German at The Johns Hopkins University, managed library coordination at the Research Libraries Group (RLG) in California, held a Council on Library Resources Management Internship at the University of Georgia, served as the Associate Director for Technical Services at the National Agricultural Library, and directed both the Cataloging Directorate and the Public Service Collections Directorate at the Library of Congress. At Cornell, she provides leadership for the 19 libraries that make up the University's library system, managing a staff of over 500 employees and 600 students. The Cornell University Library holds over 6.7 million volumes and acquires and catalogs over 100,000 titles annually.

Thomas has had a long-standing interest in information technology. She currently serves on the Executive Steering Committee of the Digital Library Federation, and she frequently speaks or writes on the topic of digital libraries. In May 1998, she was appointed a member of the New York Regents Commission on the Future of Library Services. She is a life member of ALA and serves as the chair of the Access to Information Resources Committee of the Association of Research Libraries (ARL) as well as a member of the ARL Board. She is a member of the Board of RLG and serves on advisory councils to several university libraries, including Harvard, MIT, and Washington University. Thomas earned a Ph.D. in

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

German literature from The Johns Hopkins University in 1983, writing her dissertation on the topic: "Hugo von Hofmannsthal and the Insel-Verlag: A Case Study of Author-Publisher Relations." She received her bachelor's degree from Smith College in 1970 and an MSLS from Simmons College in 1973.

Full text of paper is available

Summary: For well over a century, the catalog has served libraries and their users as a guide and index to publications collected by an institution. Charles Cutter's principles--to enable a person to find a book of which either the author, title, or subject is known; to identify all the titles held by the library on a given subject or genre, or written by a given author; or to assist in the choice of a book by edition or character--still motivate the practice of cataloging and continue to offer a framework for organization that is relevant in the world of the Internet.

The attributes of the catalog that have made it a valuable resource are desirable traits in any information management tool. Library catalogs provide those consulting them with a degree of predictability, authority, and trusted selectivity. The Library catalog user has traditionally assumed that items listed in the catalog were carefully chosen to support an institutional mission and that they were available for her inspection. Internet portals, gateways to the Web, like the catalog, offer access to a wide range of resources, but differ from the catalog in a number of ways, perhaps most significantly in that they facilitate searching and retrieval from a vast, often uncoordinated array of sites, rather than the carefully delimited sphere of the library's collections. Web information has proven much more volatile, ephemeral, and heterogeneous.

Can we re-interpret the catalog so that it can serve effectively as a portal to the Internet? Is the catalog the appropriate model for discovery and retrieval of highly dynamic, rapidly multiplying, networked documents? Until relatively recently, the catalog has been the dominant index to published literature for library users. Web portals are rapidly usurping this primacy. Libraries today are struggling as they strain to incorporate a variety of resources in diverse formats in their catalogs and to maintain centrality and relevancy in the digital world. This paper will examine the features of the catalog and their portability to the Web, and will make recommendations about the Library catalog's role in providing access to Internet resources.

Brian E.C. Schottlaender, commentator

University Librarian
Geisel Library
UCSD Library Administration
9500 Gilman Drive 0175G
La Jolla, CA 92093-0175

About the commentator: Brian E.C. Schottlaender is presently University Librarian at the University of California, San Diego. From 1993 to 1999, he worked at UCLA as Associate University Librarian for Collections and Technical Services, and served in 1998 and 1999 as Senior Associate to the University Librarian of the California Digital Library with responsibilities for primary content development. From 1984 to 1993, he worked at UCLA as, successively, Assistant Head of the Cataloging Dept. and Assistant University Librarian for Technical Services. From 1974 to 1984, he held positions at Firma Otto Harrassowitz in Wiesbaden Germany; Indiana University Libraries in Bloomington; and the University of Arizona Libraries in Tucson. A member of ALA since 1979, Schottlaender has served on the ALCT Committee on Cataloging: Description and Access since 1989 and chaired the Committee in 1992 and 1993. Since 1995, he has served as the ALA Representative to the Joint Steering Committee for Revision of AACR (JSC). In 1997 and 1998, he chaired the Program for Cooperative Cataloging. Since 1999, he has served as chair of the Pacific Rim Digital Library Alliance. Schottlaender has edited two books: *The Future of the Descriptive Cataloging Rules: Proceedings of the AACR 2000 Preconference* (1998) and *Retrospective Conversion: History, Approaches, Considerations* (1992). He has contributed articles to various professional journals, including *Rare Books and Manuscripts Librarianship* and the *Journal of Internet Cataloging*, and has spoken widely on collections, bibliographic access, and digital library issues. He received his BA degree from the University of Texas, Austin, 1974 (*ampla cum laude*) and his MLS from Indiana University in 1980, the same year he was admitted to Beta Phi Mu. In 1995, he was one of fifteen individuals selected nationally to attend the Palmer School of Library Science at Long Island University as a Senior Fellow.

Full text of commentary is available

The Catalog as Portal to the Internet

Sarah E. Thomas

Final version December 2000

INTRODUCTION

"I don't do libraries," stated an engineering student last year at an Ivy League university, pleading with his professor to absolve him from an assignment requiring him to seek information in the campus library, presumably necessitating use of the library catalog. Increasingly, even at leading institutions of higher education, one encounters not just students, but also faculty and deans who assert that they get all the information they need through the Internet. In an interview with **D-Lib Magazine** editor-in-chief and digital library scientist Bill Arms reported in the **Chronicle of Higher Education,** Florence Olsen asks Arms: (1)

> Q. Do you think, within this decade, that digital libraries will replace traditional research libraries in most disciplines?
> A. I think it may be possible to have substantial research programs without access to conventional libraries.",

Arms then provides anecdotal evidence of a colleague who meets 80% of his information needs through open source documents. Another story in The New York Times was headlined "Choosing Quick Hits Over the Card Catalog," and reported: "Even though libraries are organized and easily navigated, students prefer diving into the chaotic whirl of the Web to find information."(2)

Libraries are awash in contradictions. Gate counts are up; circulation is down. While one set of constituents eschews traditional library services, another group pushes statistics for catalog searching steadily upward. Inside the profession, librarians engage in spirited debates about their role. In the face of doubters, librarians argue that only ignorant or naïve individuals would believe that the Web could satisfy all their information needs, particularly in the scholarly community. At the same time, they energetically acquire or license digital resources.

With the addition of digital materials to the library's portfolio a debate about the role of the catalog has also developed. Should the catalog encompass all items that are considered part of a library's collection, even if those items are not physically held by the library? Should it even serve as a general gateway to the entire Web? Proponents of the catalog and of libraries believe strongly that the catalog has enduring value and that it can evolve to be a useful tool for Web access, whereas critics do not foresee any role for the library catalog as a research tool for networked information.

This paper examines the potential of the catalog to serve as a portal to the Internet. It commences with a brief overview of the development of the catalog, details the attributes and limitations of library catalogs, and defines the concept of the portal. Finally, it offers proposals to respond to the dilemma of librarians about providing access to the expanding universe of information and knowledge.

LIBRARY CATALOGS AS A PORTAL TO KNOWLEDGE

It is always humbling to learn that something you regard as a great and very contemporary problem echoes an experience from the past. Recently a small tract documenting an address to the New York State Library School in 1915 by William Warner Bishop found its way to my desk. At the time of the address, entitled **Cataloging as an Asset,** Bishop was the Superintendent of the Reading Room in the Library of Congress. Bishop's observations merit reading, even after 85 years. He notes, "the library world has seen its shifting fashions, not to say its fads of the hour. And...the striking novelties are sure to attract a good deal of attention and to get themselves much advertised."(3) Relating the change in cataloging that occurred with the Library of Congress's successful implementation of the card distribution process, he suggests that this advance had lessened the perception of the importance of cataloging, and he declares: "Catalogs and catalogers are not in the forefront of library thought. In fact, a certain impatience with them and their wares is to be detected in many quarters. Shallow folk are inclined to belittle the whole cataloging business."(4) "I think I am safe in saying," he adds, "that most students in library schools would rather do anything else than take up cataloging on graduation."(5) Bishop goes on to deplore the catalogs of booksellers, created by non-experts, and he cites approvingly the value of the permanent contributions of catalogers in the enduring description of books. In his concluding remarks he is prophetic:

> We have just begun in America, an era of huge libraries, The average size is increasing very fast. Our large libraries are getting very large. They are being run for wide constituencies on broad lines. More and more the practical American spirit is seeking for coordination and cooperation. It is by no means certain that the card form of catalog will continue indefinitely as the chief tool of library workers. It is highly probable that selected catalogs will take the place of huge general repertories. Dimly one can see the possibilities of mechanical changes and alterations, of the use of photography, instead of printer's ink, possibilities of compression or even total change of form. Certainly our present card catalogs will require intelligent direction of the highest order to make them respond to the demands of readers, to the needs of the community. Changes such as these will require an intelligent and sympathetic oversight to insure their success. The librarians who will carry them out, who will guide and mold the development of cataloging, must perforce have been experienced and trained catalogers.(6)

When Bishop wrote, almost a century ago, the catalog was undergoing a transformation, and the cataloger was under siege. Cutter's **Rules for A Dictionary Catalog** had entered the librarian's canon, but Cutter's assumption was that the catalog referenced works held by a particular institution. While his goals for the catalog - being able to find all the works by an author, to find any work by title, to find all the editions of a work, and to find all works on a given subject, with the assumption being that the catalog referenced works held by a particular institution. Union catalogs expanded the function of the catalog to serve as an index to the holdings of multiple institutions, increasing their importance in the process.

Concomitant with the emergence of the union catalog was an increase in the standardization of cataloging practice. Early in this century, The Library of Congress revolutionized catalogs through the provision of printed cards. Over 675,000 titles were available by 1915 when Bishop wrote. Consider that in 1894, William Lane, Librarian of the Boston Athenaeum, conducted a survey of university librarians on cataloging practices as part of his preparation for writing a manual on library economy. Lane stressed in his cover letter: "Please indicate what different method (if any) from that which you actually follow you would prefer if you were settling the details of your catalogue afresh unhampered by past traditions." Survey question number 5 reads: "Do you follow pretty closely any code of catalogue rules? a. The A.L.A. rules. b. Cutter's rules. c. Linderfelts translation of Dziatzko. d. Columbia College Library or Dewey' rules. e. Jewett's rules. f. British Museum. g. Bodleian Library." Although a diversity of practice still abounds in 2000, the 20th century has seen major advances in the acceptance and employment of a number of cataloging and classification tools, including the Anglo-American Cataloguing Rules, the Library of Congress Subject Headings, the Decimal Classification system, and the Library of Congress classification system.

A key catalyst for the development of more uniform cataloging was the MARC format, created in the 1960s through major leadership and innovation at the Library of Congress. MARC enabled electronic dissemination of bibliographic records and engendered networks of libraries in such entities as OCLC and the Research Libraries Group. While initially MARC's power was felt in the economies realized through copy cataloging, first of records emanating from the Library of Congress, and subsequently, from original cataloging contributed through thousands of libraries, large and small, in the last two decades, MARC's potency has increasingly derived from unleashing the potential of the large-scale union catalog for resource sharing. It is a sign of our turbulent times that during a year in which the OCLC WorldCat database grew to 41,000,000 records, with 2.2 million bibliographic records added in fiscal year 1999, a session entitled "Is MARC Dead?" held in July at the American Library Association's annual meeting attracted an overflow crowd.

Standardized bibliographic records conveyed using the MARC format also led to the rise of local systems for the management of local library holdings. The OPAC (Online Public Access Catalog) assumed rising importance, and some librarians noted with dismay that the ease and convenience of the OPAC sometimes (often) lured searchers and lulled them into a complacency with results that were incomplete. Many institutions accelerated retrospective conversion of the card catalog to ensure that historical collections and fundamental publications acquired and cataloged prior to going online did not suffer from benign neglect. Some unconventional thinkers loaded records for titles not held by their library, such as the catalog of the Center for Research Libraries, or UMI's Dissertation Abstracts, so that their clients might encounter resources, while not directly owned by their host organization, were readily accessible to them. RLG's Eureka databases and WorldCat were also considered logical extensions of the bibliographic universe available to students and researchers using a campus library.

A constant lament throughout the decades has been the insufficiency of resources to catalog all the titles acquired by libraries. Annual reports of librarians over two centuries are studded with references to accumulating backlogs. Open an annual report from any random year, turn to the section on cataloging, and almost certainly you will find a statement such as this one, drawn from the annual report of the Cornell University Libraries, 1946/47: "It is apparent from this listing of work to be done that the staff of the Catalog Department will have to be built up steadily to the point where it will be large enough to do the task assigned it. There is no other way in which the goal can be achieved. The backlog of work is very great and it will require a considerably expanded staff for a number of years to clear it up."(7) Administrators exhorted catalogers to be more productive, and in an effort to address the inexorable growth in workload as the volume of publications and acquisitions increased, catalogers, often led by the Library of Congress, introduced a number of collaborative programs to share cataloging and achieve economies. Their success in achieving enhanced productivity, though a combination of cooperative cataloging and enhanced tools, such as the cataloger's workstation, can be measured by noting that the number of catalogers employed in ARL university libraries has declined by 25% from 1990 through 1998 while the number of titles cataloged continues to rise.(8) Although some catalogers feared loss of job security if they successfully eliminated arrearages, new categories of materials to include in the catalog emerged to absorb any slack. Manuscript finding aids, guides to images, records for electronic resources, tables of contents, and other "non-book" materials competed for the attention of technical services specialists.

CATALOGS IN THE NEW MILLENNIUM

As we approach 2001, the information landscape appears to be considerably more complex than the one our predecessors populated. There is more information, the pace of change is more rapid, and the means and formats for communication are more diverse. What contribution does the catalog make in our quest to discover and retrieve knowledge? The catalog, at the level of the local institution, provides the information-seeker with bibliographic description and access to content imbued with several critical features. In addition to embodying Cutter's principles, the catalog has come to represent access to a collection deliberately shaped with a specific community in mind. This collection, by virtue of having been selected by bibliographers or some other structured process, is deemed to be of high quality. There is an implicit assumption that the works cited in the catalog are readily available for consultation. Furthermore because libraries have generally had a commitment to preserve and maintain those items they acquired, readers anticipate that a source identified today will be available in the future as well. Because they have been assembled according to standard practices and rules, by human intelligence, there is a high consistency in description, which in turn creates a high degree of predictability in results. This dependability generates an aura of trust. The user familiar with a catalog will have a high degree of confidence in the credibility of the sources contained in it. Another function of the catalog has been to link disparate materials. Until recently, the subject linkage has been chiefly among books, but in the past few years, catalogs have begun to incorporate a variety of formats, including manuscripts, visual images, audio recordings, and now, in great numbers, digital objects. Finally, although catalog searching is a seemingly free good, with host institutions assuming the cost of maintaining local catalogs and paying for the subscription costs (but not free in the case of virtual union catalogs such as RLIN or OCLC.) Even the titles and proprietary information referenced by the catalog are more often than not purchased or licensed by a library and made freely available to its users. Recent enhancements in online catalogs have improved the quality of access. Some of the features found in state-of-the art catalogs are Web access, relevance ranking, more refined keyword searching, ability to limit by date or other information, and reference linking. Thus, the functionality of the online catalog is increasing, and its proponents are convinced that it can continue to remain an essential tool for the identification and location of documents and materials of importance for researchers. Today's OPAC holds records for books and journals, films, finding aids, audio recordings, computer files, maps, and graphic images, although the preponderance of surrogates are still for monographs and printed materials. As libraries subscribe to more and more online journals, full text documents, and other digital materials, catalog records refer to publications accessible to a community through a variety of authorizations. No longer are all the citations in a catalog to holdings owned by a library; pointing to materials served remotely has become commonplace. The purity of the principle that the local catalog provides access to materials held by the host institution has become diluted slightly to accommodate items selected for community use and readily accessible, although not physically controlled by the library. On the other hand, some librarians have balked at the introduction of certain types of electronic resources into the catalog, particularly those likely to have transient URLs or which require heavy maintenance. The catalog represents stability, dependability, reliability, and quality. Its holdings have not typically been ephemeral in nature. It goes against the grain for librarians to invest in the creation of an expensive and detailed bibliographic record if the resource for which it is a surrogate, is not likely to endure for the foreseeable future, if not permanently.

Recognizing that some patrons may prefer to connect directly with online resources without being routed through the catalog, some libraries have developed separate gateways to networked resources. These gateways facilitate access to electronic materials selected by the library by providing a single point of entry, by organizing them into categories, and using metadata, often derived from their catalog records, to assist users in locating networked resources. The gateway concept appeals strongly to those for whom speedy access to online resources is a priority, and it offers many of the desirable features of the catalog, since the bibliographic control over its contents is carefully managed by librarians. Although patrons have enthusiastically adopted the gateway at many organizations, there are some flaws in its design. Of concern for the library is the expense of maintaining synchronicity between the catalog and the gateway. Although clever programs enable the cloning of bibliographic records, entries in the catalog and the Gateway are not always identical. For example, Gateway records at Cornell are organized by simple subject categories, not by LCSH, and they contain less information than the AACR2 full MARC record in the catalog from which the Gateway entry is derived.

Another issue that has burdened catalogers has been the matter of database aggregations. The phenomenon of bundling journals or databases or other electronic materials into a single resource (JSTOR, ScienceDirect), has led to a heavy workload in those institutions which have chosen to analyze each individual title in an aggregation. The dynamic nature of these aggregations, in which titles are added and dropped by the host provider on a continual basis, sometimes without notification, has significantly increased the labor entailed in adding, dropping, or modifying bibliographic records. Confoundingly, only a few suppliers of aggregations have to date seen the desirability of providing bibliographic records as a service, forcing each subscriber to repeat the effort of incorporating references to the titles they provide separately in their catalogs and/or gateways. This inefficient and wasteful situation has led to a variety of ameliorating initiatives.(9)

The Program for Cooperative Cataloging has worked with some vendors, such as EBSCO, ProQuest, CIS, and Gale, to stimulate the provision of wholesale bibliographic records to accompany subscriptions to its database aggregation.(10) These records can be loaded into a library's local system, increasing the standardization of access and saving local catalogers from the task of creating them from scratch or searching, downloading, and modifying for local use records existing in a national database This approach has had some success, but many publishers and vendors have lacked the staff expertise to create records of the quality expected by libraries. In some cases librarians have been unable to convince them that this is a service that would be worth the expense and effort of improvement.

In July 2000 OCLC put into production a service called CORC, the Cooperative Online Resource Catalog. Over 400 libraries are participating in the development of a Web-based product that uses a combination of automated tools and library collaborators to create a database of records to Web resources. Additionally, CORC includes an authority database, a pathfinder database, and a Dewey Decimal Classification Database. Users contribute URLs to the CORC database, and using automated tools, rapidly generate resource records. The system automatically suggests Dewey Decimal Classification numbers, keywords, and conducts authority checks, resulting in automatic authority control. URL maintenance is improved over its present, labor-intensive mode in local catalogs through the application of automated functions in concert with shared effort through the partners to distribute the workload. A library may export CORC records to a local catalog or gateway in either MARC or Dublin Core formats. OCLC will include CORC records in its WorldCat database.

Still another variation on the desire to manage access to Internet resources through the catalog, thereby maintaining the elements of predictability, authority, and stability of the traditional catalog, is the creation of a digital library architecture that embraces different formats and permits crossfile searching of materials cataloged, indexed, or otherwise controlled through a number of metadata schemes. Endeavor's ENCompass, currently under development, expands the view of the OPAC to enable users to direct a single query to multiple databases constructed using different encoding languages. The product is an open framework that uses metadata standards such as Dublin Core, EAD (Encoded Archival Description), and TEI (Text Encoding Initiative) to provide access to full-text resources, finding aids, and other digital objects that the ENCompass host has identified as relevant to its user community. ExLibris is developing a similar product called MetaLib. VTLS has developed a three-part approach, "Library Automation in 3V," which includes a system to handle internal library processes, a second component to support digitization, indexing, linking, and access of multimedia materials, and a third part to facilitate integration with external sources and technologies. These initiatives offer promise for the immediate future for effective access to a broader range of materials.

As noted, libraries have struggled for years to stay ahead of the rising tide of printed publications as they labored to provide bibliographic control. The Library of Congress, for example, heroically reduced its backlog of monographs over the past decade. Yet, despite some measure of success through a combination of cooperative initiatives, new technological advances, and occasional staff increases, the essential problem of cataloging or otherwise describing and analyzing the world of knowledge has remained an enormous challenge. As print indexes morphed into online databases, some voices admonished that libraries ought never to have allowed the indexing business to migrate from their domain into the commercial sector in the 1930's, since we now see the price we have to pay for access to these valuable resources escalate. The penetration of visual culture into scholarly activity necessitates improved access and more widespread dissemination of records about visual images. Other formats and materials, such as manuscripts and audio transcriptions, have ascended in importance. The interest in these materials, which have often been sequestered in special collections, has risen in part as digital technology has facilitated their visibility and accessibility. Although the backlogs in these formats (manuscripts, music, photographs, moving images, sound recordings, and maps) were even more egregious than those of books and serials, LC has sought to increase formal control over them in the past few years, and other institutions have raised the priority of their special collections as well. The numbers remain daunting, however. At one large research library, the task of converting all existing finding aids using EAD and gaining descriptive control over its entire collection of manuscripts was estimated to exceed $3 million, and since its technical services operations, using its present methodology to organize its collections, is chronically understaffed, it expected to increase this figure by a quarter of million dollars per year, taking into account the rate of new acquisitions.

During the same period that libraries have been asserting control over their backlogs of printed publications and have been shining their light on the hidden resources found in archives and special collections, the World Wide Web sprang to life. Few people had the clairvoyance to anticipate its astonishing growth and vitality. Today it registers 1.5 million new pages per day, and with a present size estimated to be in excess of 2 billion pages, it represents a major challenge to the traditional library practices. As there is mounting evidence that students, faculty, researchers, and the general public are making the Internet their information resource of the first and last resort, library values of careful selection, standardized description, and enduring access to publications are questioned as both costly and futile. A common assertion by those conversant with the Web is that library tools such as AACR and MARC won't scale in the Web environment. One digital library specialist has advanced the theory that an Internet search engine, such as Google, could replace the expensive, labor-intensive aspects of librarianship, obviating the need for catalogers, reference librarians, or selectors, or at least significantly reducing the university's dependence on them. As Bill Arms ventures in an article entitled Automated Digital Libraries:

> Quality of service in automated digital libraries will not come from replicating the procedures of classical librarianship. More likely, automated libraries will provide users with equivalent services that are fundamentally different in the way they are delivered. For example, within the foreseeable future, computer programs are unlikely to be much good at applying the Anglo-American Cataloguing Rules to monographs. But cataloguing rules are a means to an end, not the end itself. They exist to provide services to users, notably

> information discovery. Automatic methods for information discovery may not need traditional cataloging. The criterion for evaluating the new methods is whether the users find what the information they require.(11)

PORTALS AND CATALOGS

With the Web estimated to be increasing by 10 million pages weekly, the task of indexing Internet resources is clearly gargantuan, and not something that can be accomplished by even the most industrious honeybee hive of catalogers. Instead of relying on the catalog to identify and retrieve relevant web pages, users have turned instead to Web portals. The term "portal" has gained currency recently as an entry point to the web. Traffick, the Guide to Portals, www.traffick.com traces the portal's antecedent to the search engine or directory service that began to take advantage of the millions of site visits they received daily. The search engine sites recognized commercial potential by adding features that would entice repeat visits and encourage the pursuit of particular links that would advantage their partners or advertisers. In a Princeton resource published by the InSide Gartner Group, Debra Rundle offers this definition of an Internet portal:

> Internet portals originated as the librarians of the Web. The word "portal," meaning "door," has been used to characterize Web sites commonly known for offering search and navigation tools. Circa 1996, a portal was used to catalog the available content from the Internet, acting as a "hub" from which users could locate and link to desired content. Their business models consisted solely of selling advertising banner space and directing Web surfers to their desired destinations successfully (to ensure repeat business).
>
> Now portals are more than just a launching pad to content at other sites. They offer a broad array of online resources and services. Although there is no single model for what constitutes a portal, all portals offer at least five core features: Web searching, news, reference tools, access to online shopping venues and some communication capabilities (i.e., free E-mail and chat)(12)

Howard Strauss, Manager of Academic Applications at Princeton, defines a portal as a "gateway to web access" or "a hub from which users can locate all the web content they commonly need." He asserts that mandatory features of a portal include personalization, search, channels, and links, and that desirable elements are customization, role-based models, and workflow.(13)

According to Looney and Lyman, "portals gather a variety of useful information resources into a single, 'one-stop' Web page, helping the user to avoid being overwhelmed by 'infoglut' or feeling lost on the Web."(14) They estimate that 89% of the approximately 58 million Web users in the U.S. frequent portals, and they subdivide portals into categories such as the consumer portal (directory sites such as AOL, Yahoo!), community portals, which collect and organize information relating to a particular subject or interest group, vertical portals, which are often a unified site created by a particular service provider and organized on a special business topic (ETRADE), and an enterprise portal which provides a channel for intranet and external data for a corporation or university.

Portals differ significantly from library catalogs in several key ways. Like the catalog, they are built around the concept of a community, although a considerably larger body of users than the typical library catalog user. Unlike the catalog, they integrate all manner of information in their scope, rather than concentrating exclusively on "published" information. Frequently they contain a strong commercial element, with advertising prominent on their pages, and often affecting the display of search results. The search engines they employ use programs to harvest URLs and generate responses. Search queries yield large response sets, often in the thousands, and the items retrieved include duplicates, false drops, results skewed by deliberate manipulation of terms by their authors, materials of dubious heritage: in short a vast flea market of junk, collectibles, and genuine antiques. Large numbers of the URLs retrieved lead to dead ends, where the site has moved or dropped off the face of the earth or where the information has ceased to be updated. Users spend an inordinate amount of time sifting through the vast finds, often failing to locate the best resource.

The Internet portals are rife with deficiencies. They lack the very characteristics which are the virtues of the catalog. Their value, on the other hand, is lacking in the catalog. The information they access is prolific, and is often very current. With the hyperlinked aspect of the Web, it is easy to move from document to document, and the generous amount of full-text resources allows the user to mine very specific terms. There is vastly more audio and visual data available for consultation. The user can conduct her research without the inconvenience or disruption of leaving her computer, and she can readily cut and paste the results of her searches into her own documents. Result sets are ranked by relevance, and can be tailored to personal specifications. These characteristics, along with many other positive features of the Internet, excite an enthusiasm for the Internet that outweighs the deficiencies for large numbers of the population of information seekers.

Is it possible to merge the best of the portal with the strongest attributes of the library catalogs? In 1999 several library leaders began exploring a concept of a library portal in a series of structured discussions. Jerry Campbell, CIO and Dean of University Libraries, University of Southern California, a participant in these sessions, has described the proposal for a "scholars portal" in a white paper prepared for the ARL annual membership meeting in May 2000. (15) According to Campbell, the "scholars portal would promote the development of and provide access to the highest quality content on the web...." The scholars portal would foster standards and provide cross database searching. In addition, to the provision of quality content appropriate for scholarly discovery and research, it would offer affiliated services, such as reference services. The scholars portal would stand in clear opposition to the "information.coms" with their indiscriminate content and commercialized milieu.

CATALOGS AS PORTALS?

How could a library catalog serve as a portal to the Web? One thing that it could never do is function as the sole gateway to all Internet resources. Even a collaborative endeavor such as CORC could not fulfill this role, as the quantity and diversity of Web resources defy such comprehension. Even if one were to limit the candidates for control to the high quality resources contemplated as links in the "scholars portal", one should assume that the catalog would serve as only one point of access to web resources by users, who would likely have several other portals they would consult, based on their affinity groups.

Instead of striving for comprehensiveness, the goal of the catalog as portal must be to increase the ability of a community of users to meet their information needs by doing as much "one-stop shopping" as possible. By including access to web resources in the catalog, libraries would be extending to some Internet materials the same level of control that they have traditionally provided for analog formats. They would convey, through their integration in the online catalog, the credibility conferred through an affirmative selection by an intelligent being. The presence of a citation in a catalog has come to signify for the user that the source discovered is readily obtainable, that it has been chosen for its relevance to past and present foci of the community of which the searcher is a member; that the material possesses authenticity, in that the rigor of the selection process vouches in some way for its scholarly value; and that the document consulted today will be persistently available for future examination. The wrapper of the catalog conveys respectability on its contents. Readers recognize that the texts and documents referenced in the catalog represent a diversity of viewpoints, but that the universe of publications on a particular topic has been screened (or some portion of that universe) to separate out those objects which have traditionally had the greatest value for a particular constituency. In the past, those publications have had a heavy concentration of highly edited, peer-reviewed, frequently cited publications, and the virtue of the catalog for discovering materials meeting these and other unwritten standard of quality still continues. The director of Xerox's Palo Alto Research Center, John Seely Brown, parsed the difference between the web and a library, stating:

> On the Web, most information does not have an institutional warranty behind it, which really means you have to exercise much more judgment. For example, if you want to borrow a piece of code or use a fact, you'll have to assess the believability of the information. If you find something in a library, you do not have to think very hard about its believability. If you find it on the Web, you have to think pretty hard.(16)

There is a strong argument for libraries providing access to (some) Internet resources for their clients. By creating a mechanism that offers a particular subset of information seekers the ability to search citations (and more) to a pool of information that includes all formats, libraries can offer a service that increases the productivity of the searchers. They can forge a link between past knowledge, as collected and curated in library and archival repositories, and emerging ideas, as manifested in a variety of media, in a way that a search engine which restricts itself to the URL's of web pages cannot. And libraries can permit and facilitate the discovery and use of proprietary information that is not open to the independent Web searcher using a commercial portal. This licensed content may not even be located through the search engine serving that portal because of the security wall the content provider has erected to defend its property.

RECOMMENDATIONS FOR THE FUTURE

Having justified the creation of a mechanism managed by libraries to support access to Internet resources, the next question becomes: should the catalog serve as the portal to the Web? Are the tools used to build the catalog appropriate for description of Web resources? This conference will examine the flexibility of AACR2, other metadata schemes, MARC, and other standards that librarians have commonly employed to describe, categorize, and communicate information about materials held in libraries or identified by libraries as relevant to their users. These tools are good and durable instruments, and over the years I have resented comments such as "MARC costs too much to apply," or "AACR is too complicated." In themselves, these tools are not insurmountable hindrances, and in fact, they have much good to contribute to our ability to organize knowledge. Yet, at the same time, as a library administrator, I am apprehensive about applying the same standards and procedures we are using for books and journals to Internet resources.

As we move into the 21st century, we must consider reorienting ourselves and rethink the way in which we provide access to information and knowledge. Our familiar aids, such as AACR2, should be probed for the values and basic principles of organization they yield. The IFLA Functional Requirements for Bibliographic Records contribute substantially to our understanding. But we must conceive of new ways to accomplish our goals by building actively on the past while freely abandoning rules that restrain us and readily adapting new technologies. Michael Gorman has suggested a tiered approach to the description of publications that takes into account the quality of the material being described, with a progression from AACR2 through Dublin core to keyword search indices.(17) This is sensible counsel, and provides a path from the present to the future.

One of the biggest challenges facing us is the sheer volume of material that is worthy of scholars' consideration. David Levy has noted: "There is a growing awareness of attention as a highly limited resource, stemming in part from the realization that an abundance of information, good though it is in many ways, is also a tax on our attention."(18) The filtering and organizing done by libraries has the potential to serve as a labor-saving device and productivity tool for researchers in a way that is now, in the delight over the fertility of the Web for expression, only dimly appreciated by a few. But, like the enthusiasm for the automobile that propelled the acquisition of vehicles and the construction of highways but which has spawned today concern about sprawl and congestion, the Internet will seek regulation and traffic calming devices. The library catalog, or some permutation of it, can help.

To accomplish this, we must look at a number of possible changes in the way we do our business:

1. We should decisively reduce the amount of time we devote to the cataloging of books in order to reallocate the time of our bibliographic control experts to provide access to other resources, especially Internet resources, but also unique primary resources and other analog format materials.

2. In order to reduce the time spent cataloging books, we will need to investigate and implement a combination of the following :

> Using the PCC core bibliographic record (see
> www.lcweb.loc.gov/catdir/pcc/corebook.html)

Using Dublin core or a modification thereof

Accepting copy with little or no modification from other cataloging agencies, including vendors

Working with publishers, authors, and software developers to encode publications in a standard way that permits the generation of metadata from digital objects through the use of software programs

Increasing collaborative efforts nationally and globally so that publications are cataloged according to mutually acceptable standards in a timely fashion and once only.

3. To increase the functionality of the library portal/catalog, libraries need to:

Increase the scope and coverage of materials

Ensure timely access to publications

Increase the level of access from citation to full-text or increasing degrees of granularity.

Incorporate features such as reference linking, recommended titles (others who liked this title also liked:), relevance ranking, customization, and personalization that make portals so captivating

4. To ensure success, libraries shouldn't go it alone. Libraries should:

Collaborate with other libraries in a coordinated plan for the acquisition, creation of metadata, access, and preservation of materials available through portals.

Define a clear path from the local library portal to the larger scholars portal

Partner with developers of portals and search engines to share expertise in a constructive way, drawing on the best each has to contribute to the goal of effective access to information

5. Don't hide our light under a bushel. Libraries should:

Advertise the features of the discovery database, a hybrid combining some of the best features of the catalog and the portal, using local and global outlets.

Quantify the value of the laborsaving features of the portal/catalog for the community of potential consumers and for those administrating the organizations who subsidize them and stand to benefit from them

Seek new revenue (from partner portals?) to be able to expand their scope and accomplishments

Conduct and publish research documenting improved results through use of the catalog (saves time, finds more appropriate materials; titles found are accessible, etc.)

We presently lack the resources to provide access to all the information we would like to include. In addition to changing our practices to be able to expand coverage with existing funding, we should seek additional support through Congress for LC's leadership and participation, from granting agencies such as NSF and NEH to support research and pilots in the development of metadata harvesting software, crosswalking and associated access capabilities. We should seek the support of the organizations such as OCLC, RLG, and the Digital Library Federation for research in improving means of access and in fostering collaborative programs. We should work within our geographic regions, our consortia such as CIC or NERL, and other networks to accelerate the acceptance of best practices and to create linked catalogs with reinforcing document delivery and coordinated archival responsibilities. We should work within our associations and our home institutions to build a public awareness of and appreciation for the service provided by the catalog and its creators. This contribution should be documented with both the tangible contribution to members of the host institution and the intangible value of the public good the catalog represents.

The catalog can serve as a portal to the internet if the catalog is reinterpreted to be an information service which registers in a systematic arrangement those publications and documents of interest to a particular community, regardless of the form in which they appear. This discovery and access tool may exploit a variety of metadata schemes to locate materials, but it imparts unity, predictability, authority, and credibility to search results through the efforts of expert knowledge managers and the application of principles, policies, and practices of their devising. In the short term, we can expand the catalog to be more inclusive and flexible. In the near future, however, we should expect a hybrid which will adopt some of the superior features of the catalog, but which will employ an increasingly sophisticated technological infrastructure to increase the yield for information seekers. This information management tool will have evolved from the catalog and will be influenced by what we today call the portal, but it will likely have a newly coined name to represent a new concept. This "Open Sesame" service will incorporate the trusted aspects of the catalog, granting the searcher access to a realm rich with quality resources which she can easily locate and which more often than not hits the target of her needs. At the same time, the lode will yield an array of up-to-date data covering a breadth of formats and a depth of detail.

To achieve this new information medium, we will have to have the courage to risk change and to explore unfamiliar territory. Ultimately, we should figure out a new construct in which we will devote a greater proportion of our resources to providing access to materials previously left uncataloged, but which today are an important aspect of the information landscape. Accomplishing this will require a fairly dramatic shift in attention in libraries. Reallocating 10% of our cataloging resources to address this future direction may be insufficient, but even this small amount could make a noticeable difference in thinking about which attributes of the catalog have the highest priority to apply to the broader range of materials and to considering new ways of attaining the desired goals. It might be necessary to alter the way all items are processed to redirect 10% of our resources, or we might continue to treat a certain number of materials as we have, but drastically reduce the fullness of the record for others. One thing is certain: ten percent is only a beginning. We will have to organize ourselves quite differently to provide service that is meaningful, relevant, and useful for scholars and students, and if we do not do this quickly, even our worthwhile contributions will be overlooked by many whom we could aid.

The new model of information tool should draw on the wisdom of the librarian in organization, but will use the savvy of the programmer to produce the most cost-effective and accurate results possible. In its ideal realization, the successor to the library catalog will express its virtues, but will supplement them with many new features made possible through technology. The best way to accelerate the transformation of the catalog into this new entity will be to participate openly and substantively in the design of new systems into which we can transfer certain enduring values.

Notes

1. Florence Olsen, "Logging in with...William Arms: 'Open Access' is the Wave of the Information future, Scholar Says,"The Chronicle of Higher Education, Friday, August 18,2000."
2. Lori Leibovich, "Choosing Quick Hits Over the Card Catalog," The New York Times, August 10, 2000, G1, G6.
3. William Warner Bishop, Cataloging as an Asset, Baltimore: The Waverly Press,1916, p. 4.
4. Ibid., p. 7
5. Ibid., p. 18
6. Ibid., p21-22.
7. Cornell University Libraries, Annual Report 1946/47, p. 15
8. ARL:A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC, 208/209 Feb. Apr 2000, p.5
9. Karen Calhoun and Bill Kara, "Aggregation or Aggravation? Optimizing Access to Full-Text Journals, ALCTS Online Newsletter, (Spring 2000). www.ala.org/alcts_news/v11n1/index.html
10. PCC Standing Committee on Automation Task Group on Journals in Aggregator Databases, Final Report (January 2000), lcweb.loc.gov/catdir/pcc/aggfinal.html
11. (William Y. Arms, "Automated Digital libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship?" D-Lib Magazine, July/August 2000, www.dlib.org/dlib/july00/arms/07arms.html
12. www.princeton.edurundle/PrincetonPortal.htm Document #IGG-03241999-02, 24 March 1999).
13. www.cren.net/...techtalk/events/campusportals.html
14. Michael Looney and Peter Lyman, Portals in Higher education: What are they and What is their Potential, EDUCAUSE Review, July/August 2000, p.30.
15. Jerry D. Campbell, "The Case for Creating a Scholars Portal to the Web: a White Paper," prepared

for the Association of Research Libraries, April 13, 2000,www.arl.org/newsltr/211/portal.html

16. Lawrence M. Fisher, "An Interview with John Seely Brown, Strategy & Business, Issue 17, Fourth Quarter 1999, p. 93-94
17. Michael Gorman, "Metadata or Cataloging? A False Choice." Journal of Internet Cataloging, v.2, no. 1 1999, p. 5-22
18. David Levy, "I Read the News Today Oh Boy: Reading and Attention in Digital Libraries, Proceedings of the 2nd ACM international conference on digital libraries, July 23 - 26, 1997, Philadelphia, PA USA" p. 202-211 (p. 202)

Library of Congress
December 21, 2000
Comments: lcweb@loc.gov

"The Catalog As Portal To the Internet"
by Sarah E. Thomas

commentary by Brian E. C. Schottlaender

Final version

Ms. Thomas' good and thoughtful paper, and its assorted "modest" proposals, challenge us to ask ourselves eight questions-some of which are embodied in her closing recommendations, others of which are not. I shall, in this commentary, respond briefly to each in turn.

Q: Should libraries create and manage a mechanism to support access to Internet resources?
A: YES.

In fact, libraries already do manage many such mechanisms. We have all no doubt heard of, and perhaps even used, Cornell University's Gateway, the University of California San Diego's Sage, and the University of Wisconsin's Scout-to name a few. The problem, as I know Ms. Thomas knows, is that libraries have not created "a [single] mechanism." We have created several. As a consequence, not only must we manage several, but our clients must navigate several.

Q: Are the tools used to build the catalog appropriate for description of Web resources?
A: YES AND NO.

The tools used to build the catalog will work (well enough) to describe Web resources, but it may not always be appropriate (or desirable) to use (all of) them all of the time.

- First, library catalogs comprise various flavors of AACR and MARC, various subject heading and classification schemes, and various authority control and records management processes-individually and in combination. "The catalog," as Tom Delsey intimates in his paper, is, thus, not a monolithic construct.
- Second, library catalogs have never included all materials "privileged" into a library's collections, and not just as a consequence of cataloging backlogs. In fact, the "tools used to build catalogs" have never lent themselves to describing all of our collections. Access to journal articles, for instance, has long been managed using the tools by which A & I databases are now created. Individual items in archival collections, by way of additional example, have long been managed via finding aids (if that).
- Finally, a goodly number of Web resources have no analogs amongst the material types long described in library catalogs: neither physical/structural analogs, nor intellectual analogs. These, thus, are unlikely to lend themselves to description with tools used to build library catalogs.

Q: Should the catalog serve as the portal to the Web?
A: NO.

It can't, it shouldn't, and it doesn't need to. In fact, catalogs and portals are "metadata constellations" which, when integrated-as, for the most part, they presently are not-make up, along with other such constellations, the "universe of access." To ask catalogs to serve as portals to the Web is asking too much of them, just as asking portals to serve as catalogs of "the non-Web" is asking too much of them. I do not believe that I disagree with Ms. Thomas in taking this position inasmuch as her own argument speaks of "reinterpreting" the catalog, and of imagining a "new information medium" that is a hybrid combining some of the best features of the catalog and the portal. She is, thus, no longer talking about the catalog.

Q: Should libraries "decisively reduce the amount of time we devote to the cataloging of books in order to reallocate the time of our bibliographic control experts to provide access to other resources, especially

Internet resources. . . ?"
A: NO.

Ms. Thomas' "InfoGlut" slide describes an international publishing environment in which book production is averaging one million volumes published annually. In that sort of environment, libraries will not be in a position to "decisively reduce" the amount of time devoted to book cataloging, at least not until such time as it has been demonstrated that:

- the amount of time devoted to the cataloging of books is disproportionate to the quantities of them acquired for our collections, and that
- the amount of time devoted to their cataloging is disproportionate to the interest shown in them by our clienteles.

Further, Ms. Thomas' recommendation unnecessarily and undesirably dichotomizes between books and all other information resources. Libraries were uncomfortable with the "access vs. ownership" dichotomy; Michael Gorman has suggested that the "cataloging vs. metadata" dichotomy is a false one1; this one is no better.

Finally-and ironically-Internet resources lend themselves to automated processing ("cataloging") in ways that [printed] "books" do not. And yet, we've barely begun to explore how best to take advantage of these automated processing capabilities. Better, at this point, we should continue to explore that avenue than go down that of reducing our cataloging commitment to those materials which do not lend themselves to such processing.

Q: Should libraries investigate and implement a combination of the following:

a. using the PCC core bibliographic record;
b. using Dublin core or a modification thereof;
c. accepting copy with little or no modification from other cataloging agencies, including vendors;
d. working with publishers, authors, and software developers to encode publications in a standard way that permits the generation of metadata from digital objects through the use of software programs;
e. increasing collaborative efforts nationally and globally so that publications are cataloged according to mutually acceptable standards in a timely fashion and once only?

A: YES.

It is highly desirable that libraries pursue any and all of these strategies, although not to "reduce the time spent cataloging books" specifically, but, rather, to maximize the time spent cataloging generally. Of these, (a), (c), and (e) in combination have the most promise, while long-term investment in (d) may or may not yield a long-term dividend.

Q: Should libraries increase the functionality of their catalogs/portals by:

a. increasing the scope and coverage of materials;
b. ensuring timely access to publications;
c. increasing the level of access from citation to full-text or increasing degrees of granularity;
d. incorporating features such as reference linking, recommended titles (others who liked this title also liked:), relevance ranking, customization, and personalization that "make portals so captivating?"

A: YES AND NO.

(a) and (b) above are non-controversial because libraries should always have provided timely access to materials of sufficient scope and coverage to meet the needs of their clientele. Unfortunately-and this is no doubt part of Ms. Thomas' point-"libraries should always have" does not mean "libraries have always." (c) above is more debatable because what level of access is appropriate when and to whom is itself (or are themselves) debatable.

Finally, (d) above is especially debatable because if one polled the 100+ attendees at the LC Bicentennial Conference on what features they think "make portals so captivating," one would probably not only get 100+ different answers, but half the attendees would probably disagree with the other half. One person's "captivating" is another's annoying, erroneous (cf. the amusing discussion in Thomas Mann's paper of the "recommended title" feature found on a number of portal sites), or presumptuous. Ms. Thomas' slide describing corporate portals as "competing for the eyeballs of their [i.e., the corporations'] employees" is symptomatic. Would that portals sought to compete for the minds of those using them, rather than our eyes!

Q: Should libraries not "go it alone," but instead:

a. collaborate with other libraries in a coordinated plan for the acquisition, creation of metadata, access, and preservation of materials available through portals;
b. define a clear path from the local library portal to the larger scholars portal;
c. partner with developers of portals and search engines to share expertise in a constructive way, drawing on the best each has to contribute to goal of effective access to information?
A: YES.

Collaboration and partnering both facilitate standards development and implementation and reduce unwanted redundancy. Libraries have long been fairly good at cooperating with each other. As noted by Priscilla Caplan in her paper, however, we've not been terribly good at collaborating with those outside our own community. Just as, for example, the librarians who developed the EAD DTD did so in concert with DynaWeb's software developers, so too will it behoove libraries in general to work with Intel and the like to pursue the development and refinement of Internet discovery mechanisms. It is worth noting that Ms. Thomas' "Manage the Knowledge of Thousands" slide depicts not a confident young librarian striding into the future, but, rather, a Lotus employee doing so!

Q: Should libraries bring their "light" out from under the "bushel basket" by:

a. advertising the features of the discovery database, a hybrid combining some of the best features of the catalog and the portal;
b. quantifying the value of the laborsaving features of the portal/catalog for its clientele;
c. seeking new revenue (from partner portals?) in order to expand their scope and accomplishments;
d. conducting and publishing research that documents improved results through use of the catalog (saves time, finds more appropriate materials; titles found are accessible, etc.)?

A: YES AND NO.

Researching, quantifying, and publicizing ("advertising") the features of the "discovery database" and its time and scope implications are good ideas. Whether the discovery database is a physical construct-as Ms. Thomas' characterizing it as "a hybrid combining some of the best features of the catalog and the portal" makes it sound-or a logical construct which amalgamates into a single view (though, as suggested in Thomas Mann's paper, perhaps not a "seamless" view) the range of relevant resources offered up by each remains to be seen.

Notes 1.)Michael Gorman. "Metadata or Cataloging?: A False Choice." Journal of Internet Cataloging 2(1): 5-22.

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Tom Delsey

Director General
Corporate Policy and Communications
National Library of Canada
395 Wellington St.
Ottawa, Canada K1A, ON4

The Library Catalogue in a Networked Environment

About the presenter:

Tom Delsey is presently Director General, Corporate Policy and Communications, at the National Library of Canada (NLC). During his twenty some years with NLC, he has been Chief of the Canadian MARC Office, Assistant Director for Standards, Director of the Cataloguing Branch, Director of the Acquisitions and Bibliographic Services Branch, and Director of Policy and Planning. Over that same period of time, Delsey has been an active participant in various Canadian and U.S. committees, including the Canadian Committee on Cataloguing, the Canadian Library Association's Copyright Committee, MARBI, and the CONSER Executive Committee. Internationally, he has been involved in several committees and working groups within the International Federation of Library Associations (IFLA), International Standards Organization (ISO), and ISDS. In recent years he has served as a consultant for the IFLA Study on Functional Requirements for Bibliographic Records, and most recently he has completed a study for the Joint Steering Committee for Revision of AACR (JSC), producing a schematic model of the cataloguing code's logical structure. He has published a number of papers on bibliographic standards and the application of technology to bibliographic control. Delsey has a bachelor's degree in English literature from McMaster University, an MLS from the University of Western Ontario, and a Ph.D. in English and American language and literature from Harvard University.

Cataloging Directorate Home Page

Library of Congress Home Page

Full text of paper is available

Summary:

For the past four decades the development of the library catalogue has been inextricably linked to advances in digital technology. In the sixties, libraries began experimenting with the use of digital technology to support catalogue production through the capture, formatting, and output of bibliographic data. In the seventies, software developers introduced a wide array of systems to support online access to library catalogues. In the eighties, the development and implementation of open standards added a new dimension to the networking of online catalogues. In the nineties, the development of Web technology enabled libraries for the first time to link records in their online catalogues directly to the digital resources they describe.

The evolution of the library catalogue from a manual to a digital form has had a significant impact on the interface between the catalogue and the user, and in a number of fundamental ways has altered the way in which catalogue data is accessed. Likewise, the migration of the catalogue from a local to a networked environment has had a significant impact on the interface between the individual library catalogue and other catalogues and bibliographic databases accessed through the network. Potentially, even more significant for the library catalogue is the direct linking to digital resources that is made possible through the World Wide Web and the impact of this new technology on the interface between the catalogue and the resources the catalogue describes.

This paper provides an overview of how technology has changed the relationships between the library catalogue, the catalogue user, alternative sources of bibliographic data, and the resources described in the catalogue. It looks--from a technical perspective--at what those changes mean for the way we support various interfaces to the catalogue, and it highlights changes in approach that will be needed in order to maintain and enhance the effectiveness of those interfaces in an evolving networked environment.

Jennifer Trant, commentator

Executive Director
Art Museum Image Consortium
2008 Murray Ave., Suite D
Pittsburgh, PA 15217

About the commentator: Jennifer Trant is the Executive Director of the Art Museum Image Consortium (AMICO) http://www.amico.org. She also serves as a Partner in Archives & Museum Informatics, and Editor-in-Chief of *Archives and Museum Informatics: the Cultural Heritage*

Informatics Quarterly from Kluwer Academic Publishers. She is co-chair of "Museums and the Web" http://www.archimuse.com/mw2000/ and "ichim2001" in Milan, Italy http://www.archimuse.com/ichim2001/, and is on the program committee of the ACM Digital Libraries 2000 conference, and the Board of the Media and Technology Committee of the American Association of Museums.

Prior to joining Archives & Museum Informatics in 1997, Jennifer Trant was responsible for Collections and Standards Development at the Arts and Humanities Data Service, King's College, London, England. As Director of Arts Information Management, she consulted regarding the application of technology to the mission of art galleries and museums. Clients included the now closed Getty Information Institute for whom she managed the Imaging Initiative and directed the activities of the Museum Educational Site Licensing Project (MESL). She also prepared the report of the Art Information Task Force (AITF), entitled *Categories for the Description of Works of Art for the College Art Association and the Getty.*

Full text of commentary is available

Library of Congress
January 02, 2001
Comments: lcweb@loc.gov

The Library Catalogue in a Networked Environment

Tom Delsey
Director General
Corporate Policy and Communications
National Library of Canada
395 Wellington St.
Ottawa, Canada K1A, ON4

Final version

With the migration of the library catalogue to a networked environment there have been a number of significant technological changes in the way cataloguing data is accessed and utilized. As the OPAC has been supplemented by other technologies-search and retrieval protocols, browsers, search engines, and resolution services-the interfaces between the catalogue and the user, between the catalogue and the library collection, and between the catalogue and other sources of data on the network have become increasingly complex, both in the way they are structured and in the level of functionality and interoperability that they support. To understand more fully the way the catalogue functions in a networked environment, and how its functionality can be optimized, it is important to view the catalogue not simply as a data store, but more broadly as the interaction between that data store and a growing range of networked applications that interface with the catalogue.

This paper is intended to do just two things. The first is to sketch out in broad terms the impact that technological change over the past few decades has had on a number of key interfaces to the library catalogue. The second is to highlight, again in fairly broad terms, certain aspects of those interfaces that will need to be analyzed more closely as we endeavour to make the library catalogue a more effective tool for accessing networked resources. My purpose is simply to help establish a frame of reference or context for some of the more specific needs, challenges, and potential solutions that will be addressed in greater detail in the dozen or so papers that follow.

The Impact of Technology on the Interfaces

There are two interfaces that are absolutely integral to the functioning of the library catalogue: the interface with the user and the interface with the resources described in the catalogue. It is through those interfaces that the catalogue fulfills its primary function of facilitating access to the library's collection. There are, however, two other key interfaces that come into play as a means of supplementing the functionality of the catalogue. One is the interface between the catalogue and the tools produced by abstracting and indexing services. The other is the interface between the local catalogue and the union catalogue. The A&I interface serves to supplement the level of content analysis that is provided by the catalogue itself. The union catalogue interface has served to supplement the reach of the catalogue, facilitating access to the library's collection for a wider group of users than the library's direct clientele.

In the transition of the library catalogue from its card format to the OPAC, and the subsequent migration of the OPAC to the Internet and the Web, there have been significant impacts on all four of the interfaces to the catalogue. A brief overview of the changes that have occurred with respect to each of the interfaces will serve to highlight how significant some of those changes have been and what kind of challenges we face in adapting the interfaces to a new technological environment.

The User Interface

The most obvious impact of online technology on the user interface with the library catalogue has been the extension of access. A machine-readable database of catalogue records, by effectively eliminating the physical constraints associated with the card catalogue, brings with it the potential to give the user access to virtually any element of data within the catalogue. With online access to the catalogue, the traditional access points provided in the card catalogue have been supplemented through the indexing of a variety of additional data fields, extending the scope of the user's searching capability significantly. Computer indexing has also served to extend the functionality of the individual access point. Permutation of conventionally structured headings makes it possible to search the catalogue not only using the lead element in such headings, but using any sub-element of the heading, whether it be the name of a corporate body recorded as a sub-heading, or a form subdivision used in a subject heading. Keyword indexing has extended the search capability even further. And the addition of Boolean search functions has given the user the capability of extending or narrowing a search in ways that simply were not possible with the fixed structure of the card catalogue.

Online technology has also had a significant impact on the way catalogue data is displayed. The conventional "unit record" display of the card catalogue has been displaced by what is typically a graduated display starting with one or more "results set" screens, from which the user is given several options for the display of individual records, ranging from some form of brief record, through a full record in a conventional catalogue entry format, to a display of the record with all its MARC coding. In addition, the online catalogue offers the user a range of options for sub-arranging the records that form the results set for the search, as well as the capability of combining results sets.

These new capabilities for both search and display of catalogue data have had the effect of substantially altering the underlying structure of the library catalogue. The structure of the card catalogue was effectively pre-determined by the form in which headings and references were cast, by the format of the "unit record," and by the conventions used for filing individual entries within an established sequence. The standardization of cataloguing rules, card formats, and filing rules served to establish a uniform structure for the card catalogue that was in all essential respects consistent from one library to another. With the introduction of the online catalogue all that was changed. The opportunities that the technology provided for extending access to the data stored in the catalogue and expanding the range of display options led to innovations in the design of the user interface that have not only altered the nature of the catalogue as a search and retrieval tool, but have effectively displaced what had been a common structure with a multiplicity of structures.

From the user's perspective, this migration of the library catalogue to an online environment meant that the interface with the catalogue had to be re-learned. What had been a relatively simple tool, the structure of which could be understood more or less intuitively, and the use of which required little technical skill, had been displaced by a tool that was considerably more complex in its design and utilized a new technology that in itself required the user to develop a new skill set.

The second stage in the migration of the catalogue-from what was effectively a "local" online environment to a fully networked environment-has brought with it a new set of challenges. The innovation that was sparked with the introduction of online technology, and the wide-ranging variations in the design of database structures, indexing methods, and systems functionality that have ensued, have made the design and implementation of user interfaces in a networked environment all the more complex. In the "local" online environment, the user interface was designed to function within the context of a particular database structure, a specific set of indexed data elements, a defined set of processing capabilities, and an established range of functionality at the desktop. In a networked environment the potential for variability within those interface dependencies is virtually infinite.

The Resource Interface

The migration of the catalogue to an online environment has until recently had a relatively minor impact on the interface between the catalogue and the resources it describes. The reason, of course, is that prior to the more recent development of Internet and Web technologies, the interface between the catalogue and the collection of resources it described has had to bridge what are effectively two separate environments. As long as the library collection itself remained exclusively a physical collection stored on shelves and in cabinets on the library premises, any direct connection between the catalogue and the collection was impossible. As a result, the resource interface continued to function in the same way as it had prior to the computerization of the catalogue. That is to say that the interface continued to be dependent on a locally assigned data element in the form of a call number or shelf number appended to the catalogue record that served to identify the location of the item described within the collection as a whole, but otherwise provided no direct link to the resource.

More recently, as libraries have begun to add networked resources to their collections, it has become feasible to create a direct link between the catalogue record and the resource described. To this point, most of those links have been established by means of a Uniform Resource Locator (URL) that functions, through Internet protocols, as an accessible address for the resource. The link is effective, of course, only as long as the address remains valid. And therein lies the first challenge.

To be effective in supporting the link from the catalogue to the resource described, the identifier on which that link is based must remain valid over time. As library collections become increasingly "virtualized," maintaining the validity of the identifier becomes increasingly problematic. For resources that are stored on a server under the direct control of the library, the continuing validity of the identifier can be achieved through effective management of the library's own repertoire of URLs. But for resources that are stored on servers outside the library's control, the continuing validity of the link is entirely dependent on the data management practices of the host organization. And that is equally true for any identifier that labels itself as a "persistent" identifier (such as a PURL or a DOI) as it is for a simple URL. If the host organization fails to maintain the link between the resource and any identifiers that have been used to support the link to that resource over time, those identifiers simply will not work, regardless of whether they purport to be persistent or not.

A second challenge to the resource interface arises from changes in the nature of ownership in the collection that are the result of extending the library collection to encompass networked resources. Traditionally libraries have served their users by making items in their collections available for onsite use, for loan, or under certain circumstances, by making a copy of a portion of an item's content. Such uses have been predicated on the library having physical ownership of the copies in its collection, having the right to lend such copies, and having the right through exceptions in copyright law to make copies in accordance with specific criteria. With the introduction of digital resources, and in particular networked resources, the proprietary relationship has changed, and the library's entitlement to make those resources available for use is increasingly governed by contractual licence. The application of copyright in a digital environment, as reflected in recent judicial decisions and amendments to copyright law, is also significant.

From a technical perspective, the increased complexity associated with access rights to networked resources will have a significant impact on the interface between those resources and the library catalogue. The interface will have to function as more than a simple link from the catalogue record to the resource described. The resource interface may in fact have to be reconceptualized to function in tandem with the authentication procedures in the user interface to support the administration of terms and conditions embodied in contractual licences and perhaps even to monitor uses permitted under copyright law.

The Abstract/Index Interface

The tools produced by abstracting and indexing services have always been used by libraries as a pragmatic means of extending bibliographic access to the contents of their collections. Such tools provide a level of content analysis for journal literature, conference proceedings, compilations, and anthologies that libraries are rarely able to provide through the catalogue itself.

The application of online technology both to the library catalogue and to abstracting and indexing tools has served not only to improve access but also to increase the efficiency of the interfaces between the abstracting and indexing tools and the library catalogue in a number of ways. The most notable impact on the interface has been the effective integration of access to data describing articles, papers, etc. contained in the journals and conference proceedings held by an individual library with access to data recorded at the serial or monographic level in the library's catalogue. Integrated access has been made possible, in large part, through the widespread use of standard identifiers such as ISSNs and ISBNs both in the citations that are created for the abstracting and indexing tools and in the monographic and serial records created for library catalogues. Additional support for integration has come from the systematic enhancement of serial records through initiatives such as the CONSER A&I Project to include structured data fields identifying specific tools in which the contents of the serial described in the catalogue record are indexed. With those kinds of data links in place it has been possible for libraries to extract customized subsets of abstracting and indexing data relevant to their individual collections and to use their local OPAC software to provide access to their holdings of serials and conference proceedings at an analytical level.

As abstracting and indexing databases move to a networked environment, and as the scope of A&I services is extended increasingly to coverage of electronic journals and other networked resources, the relationship between the user, the catalogue, the A&I database, and the electronic resources that both the catalogue and the A&I databases provide access to has the potential to be substantially altered. Standard search and retrieval protocols open up the possibility of providing another alternative to the OPAC as a means of accessing analytical data derived from multiple A&I sources through a single search. In addition, where the journal or other source referenced in a citation is in electronic form, accessible through the Internet, networking technology makes it possible for the creators of A&I databases to link their citation data directly to the electronic article or document cited. Technically speaking, routing the output from an A&I search through the library catalogue in order to provide the user with a copy of the article or document cited is no longer a pre-requisite.

What remains, however, is a need, at least in certain cases, for the library to serve as an intermediary in validating the user's access rights to the electronic resource. If access to the resource is restricted to licensed subscribers, and the user is accessing the resource as a user of a particular library, it will be necessary to verify that the user is entitled to access under the library's licence. Making that connection between the user and the library thus becomes a legal pre-requisite, and introduces added complexity to the relationship between the user, the A&I database, and the electronic resource. In effect, it becomes necessary in a networked environment to re-establish an interface between the A&I database and the library catalogue that will support user access to electronic resources in the library's collection that is not entirely dissimilar in function to the interface between A&I data and catalogue data that has been established to operate at the local level through OPAC software.

The Union Catalogue Interface

The union catalogue has traditionally functioned as a means both of extending the reach of the local catalogue and of supplementing its scope. Holdings reported to union catalogues have served to make the reporting library's collection accessible to a wider group of users than would normally be served by the local catalogue. In turn, having access to union catalogues has served as a means of meeting user needs that cannot be fulfilled through the local catalogue.

With the application of online technology to both the local catalogue and the union catalogue, the interface between the two began to change in a number of technical respects, but its basic nature remained largely unaltered. Holdings that had previously been reported in the form of cards, printed lists, or microform began to be reported in machine-readable form, first on tape and subsequently through file transfer protocols. Editing and de-duplication processes were automated to the extent possible, but continued to be supplemented through manual follow-up procedures.

Initially the introduction of online technology in fact had less impact on the interface between the local catalogue and the union catalogue than it had on the user interface to the union catalogue itself. The new search capabilities that were available through online technology served to make the union catalogue, for the first time, and in most respects, as effective an access tool as the local catalogue. Prior to computerization, the union catalogue had functioned in a much more limited way than the local catalogue, largely because the physical constraints of the card catalogue and the labour required to compile and edit the catalogue made its implementation as anything other than a single entry catalogue impractical. But once the card catalogue was replaced with a machine-readable database it became possible to exploit the power of online technology as fully with the union catalogue as with the local catalogue.

With the introduction of the Internet, however, there has emerged an alternative means of extending the reach and supplementing the scope of the local catalogue. With networked support for search and retrieval protocols such as Z39.50, the union catalogue has been reconceived as the virtual union catalogue. The potential advantages to be gained through implementation of a virtual union catalogue are considerable-elimination of the costs associated with compiling and maintaining a separate union catalogue database, more flexibility in establishing the scope of libraries to be included in a union catalogue search, more timely "reporting" of new accessions and withdrawals, and a seamless interface to data on the current availability of an item targetted for loan. What remains to be seen, though, is whether implementation of the supporting protocols can be managed in such a way as to realize those potential benefits across a critical mass of library systems. The other key challenge for the virtual union catalogue is to find a means of achieving "on the fly" what has been achieved in the conventional union catalogue through systematic quality control and de-duplication procedures.

Areas of Focus for Future Development

Addressing the challenges posed by the migration of the catalogue to a networked environment is going to require the involvement of the library community in a multiplicity of assessment and development initiatives. The range of issues raised as a result of technological change is broad and complex. Each of the interfaces to the catalogue is affected in different ways, and new interdependencies have emerged between and among the interfaces.

Stepping back and looking at the impacts in the aggregate, there would appear to be three broad areas in which future development needs to be focused. The first centres on the data itself. If the catalogue is to function as an effective tool for facilitating access networked resources, we need to ensure that the data recorded in the catalogue is adaptable to the description of those resources and that it is adequate to support the various applications that will draw on it. The second relates to the functionality supported by the interfaces. Again, if the interfaces are to support a wider range of functions and to operate within in a more complex architecture, we will need to ensure that the requirements and interdependencies are fully understood. Thirdly there is the issue of strategic positioning of the catalogue. This new environment requires extensive rethinking not just of how the technology can be exploited, but also of how the catalogue, and by extension the library itself, can be repositioned to meet the needs of its users.

Reassessing Data Requirements and Conventions

In comparison with the scope of technological change that has occurred with the migration of the library catalogue to a networked environment, there has been relatively little change to date in the bibliographic conventions used by libraries to compile data for those catalogues. Cataloguing rules have been updated in an incremental way over the past three decades to accommodate the description of an evolving repertoire of information carriers, and MARC formats have been enhanced to some extent to respond to current technical developments in data management, but the rules and formats remain strongly rooted in earlier technologies, and there is a growing gap between the conventions reflected in cataloguing rules and formats and the technological environment within which the catalogue currently operates.

As the nature of the resources available through the Internet and the World Wide Web evolves, and as the user's approach to resource discovery changes in response to features built into browsers, search engines, and other tools available on the network, it is essential for libraries to take a closer look at the data used in resource discovery and the way in which it is used. That process might usefully start with a review of the matrices developed for the Functional Requirements for Bibliographic Records that mapped attributes and relationships associated with the various entities reflected in catalogue records to the generic user tasks-find, identify, select, and obtain.[1] What needs to be determined is whether there are attributes or relationships associated with networked electronic resources (at either the logical or the data element level) that have significant value to the user engaged in resource discovery that are not currently reflected in catalogue records. That review needs to focus not only on data required to assist the user in finding resources in response to a search query, but also on data required to assist the user in assessing the relevance of the resources found and determining the usability of the resource from a technical perspective as well.

At a deeper level there is a need to revisit the cataloguing conventions that are currently used to describe resources in library collections and to determine standard access points and citation forms for the works contained in those resources. The analysis of the Anglo-American Cataloguing Rules that was undertaken recently for the Joint Steering Committee revealed a number of structural issues relating to the internal logic of the cataloguing code that need to be addressed if AACR is to serve as an effective tool for cataloguing digital resources.[2] Embedded in the logic of the code there are implicit assumptions derived from the traditional view of the resource as a physical object that make the application of the rules to networked resources highly problematic. A key issue to be addressed is how to adapt cataloguing data conventions to accommodate the description of resources whose content is not fixed in the way it was in non-digital media and is so susceptible to transparent alteration and extension.

On another front, data requirements for support of the interface between the library catalogue and the resources described in the catalogue need to be reassessed in the context of the direct linking to networked resources that is now possible. Libraries need to evaluate the relative strengths of the various identifiers that might be used to support the link from the catalogue record to the resource and determine how to achieve the persistency that is required of that link. Over and above the link itself there is a need to determine data requirements related to access rights. Although data relating to the "purchase" of a resource has not normally been recorded in the catalogue per se, logically such data, being both library-specific and resource-specific, needs at least to be linked to the data the library maintains in the catalogue to identify the resource, and the resource interface needs to draw on and link both types of data.

By extension, data relating to access rights acquired by the library will come into play as well in the union catalogue interface. With the addition of networked electronic resources to library collections there will be a need to indicate whether access to a particular resource is restricted to the library's direct users, or whether access through an arrangement analogous to interlibrary loan is possible, and if so, under what conditions. In that context there may be a need for additional data relating to access rights acquired, for example, through a consortium licence, that would be relevant to a user conducting a protocol supported search of a virtual union catalogue.

Re-examining the Interfaces

As noted earlier, to understand the way the library catalogue functions in a networked environment the catalogue needs to be viewed not simply as a database but more broadly as the interaction between the database and the applications that interface with it. To understand how the catalogue's functionality can be optimized in a networked environment, it is necessary, therefore, to re-examine not just data requirements but the functional requirements supported by the interfaces as well.

Looking, for example, at the changes that have occurred in the transition from the manual catalogue to the OPAC, and in turn from the OPAC to the networked catalogue, it is clear that the functionality supported by the user interface has changed significantly, with increased search capabilities and greater flexibility of display. However, when comparing the support that a typical client application offers for organizing a display of multiple records for various versions and editions of the same work with the logical sequencing of those same records in a card catalogue, it is not always so clear that the functional support provided by the online interface is an improvement over its predecessor.[3]

A similar observation can be made regarding the union catalogue interface. In a pre-networked environment, the usability of the union catalogue was heavily dependent on the editing and de-duplication processes that were an extension of the "reporting" mechanism, and in effect part of the interface between the local catalogue and the union catalogue. With the implementation of the Z39.50 protocol and the development of the virtual union catalogue, those editing and de-duplication processes have been relocated, as it were, to the client application, where they have to be executed "on the fly" with each results set. Current implementations of Z39.50 client software in fact seldom provide that level of functionality. Add to that the shortcomings of client applications in supporting logically organized displays of results sets, and it should be fairly evident that further development is needed to bring the interface with the virtual union catalogue up to par.[4]

A re-examination of the functionality incorporated into Z39.50 client software might be extended further to include an assessment of the potential for such applications to support a networked interface between the catalogue and abstracting and indexing databases. In a networked environment it is technically feasible to achieve through a Z39.50 or other protocol based interface what in a local OPAC environment could only be achieved by maintaining on a local server copies of records derived from commercially produced abstracting and indexing databases, pre-selected to correspond to the library's serials holdings. If protocol supported client software is to serve that purpose effectively, however, it will be necessary first to establish a broadly based framework for interoperability between client applications at the library end of the interface and target applications at the A&I end. In practical terms, the most effective means of developing such a framework would likely be through an extension of the work that is currently underway with the development of the Bath Profile.[5] Key elements in the interface that would need to be examined are the identifiers, both at the article level and at the serial level, that are now being used in A&I database citations for networked resources.

With the migration of the resource interface to a networked environment functionality issues of another kind emerge. As noted earlier, prior to the introduction of Internet and Web technologies there was in effect no technical means of fully supporting the interface between the catalogue and the collection of resources it described. OPAC technology could be used to generate a call slip (or its equivalent as a screen display), but from there it was left to the user or a library employee to manually retrieve the item from the stacks. Now with the capability of linking directly from the catalogue record to the resource described (at least in the case of networked resources) a new dimension of functionality is brought into play. The resource interface becomes in effect a resolution service, or at least the front end to a resolution service.

Technically, resolution in a networked environment is fairly straightforward. What is more complicated, however, is designing mechanisms that will facilitate resolution that is consistent with proprietary and contractual arrangements associated with a particular resource. It cannot be assumed that resolution from the description of a resource in a library catalogue directly to the originator of the resource will invariably be the preferred route. There will be cases where the library requires a connection to be made indirectly via a supplier or aggregator who manages the library's licence for access to the resource. There will also be cases where a connection to an archived version of the resource housed on a server maintained by the library itself is required. What needs to be examined more closely is whether the mechanisms embedded in the network per se will be sufficient to support the kind of selective routing that a library may require, or whether that kind of functionality needs to be built into the library's end of the resource interface.

Repositioning the Catalogue

Optimizing the performance of the library catalogue in a networked environment will clearly require a significant level of effort in the technical redesign of data structures and applications. Much of that work will have to be carried out at an international level, and will involve a significant degree of cross-sector cooperation. But optimizing performance through the exploitation of network technologies is not all that will be required to position the library catalogue strategically within this new environment.

The technology that supports the direct linking of catalogue records to the electronic resources they describe is also being used to support links to those same resources from a wide range of network browsing services, Web directories, indexing tools, and publishers' databases. The same technology also supports direct links from references and citations embedded in an electronic document to the resources referenced. Likewise, the technology that supports the horizontal extension of the local catalogue through the virtual union catalogue or through a networked interface between the catalogue and an A&I database is being used in other sectors as well to extend local functionality for resource discovery across multiple sources of data. What all this means, of course, is that the library catalogue functions as just one of many access paths available to the user in search of electronic resources on the network.

Positioning the library catalogue as a primary access mechanism within this environment will require a strategic focus not only on the technologies that are being broadly deployed throughout the network, but also on those aspects of the catalogue that are integral to its design and serve to differentiate it from other access mechanisms. One such element is the cataloguing process. The value of the catalogue as an access mechanism is derived in large measure from the quality control inherent in the data creation process-in the consistent application of descriptive standards, the control of name and title access points through authority files, the development of subject thesauri and classification schemes, and the standardization of formats and coding for machine-readable records. Added value is derived as well from the wide-scale adherence to cataloguing standards within the library sector, which means that in the aggregate library catalogues have the potential to function effectively as an integrated access mechanism to an enormous store of resources.

Equally important from a strategic perspective is the fact that the library catalogue functions as a guide to a collection of resources professionally searched, selected and maintained for the purpose of supporting the research and information needs of a defined community of users. With the exponential growth that characterizes the Internet, the selectivity and pre-determination of relevance that are reflected implicitly in the library catalogue take on even greater value. The library catalogue also differs from many of the newer access mechanisms on the network in that it is has a retrospective as well as a current dimension to its design and function. The fact that as an access tool the library catalogue, like the library collection itself, has an archival function is of critical importance in a networked environment so widely evanescent in nature.

Setting the agenda for the adaptation and development of the library catalogue to function more effectively in a networked environment is in itself a challenging task. Clearly there is a need to exploit new technologies as fully as possible. Likewise, there is an increasing need to factor cross-domain interoperability into the equation. But there is also a need to retain and enhance to the extent possible those features of the catalogue that have served over time to make it an effective tool for its users and that give it the potential to outperform other resource discovery tools in this new environment.

Notes

1. IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report.* Munchen: K.G. Saur, 1998.
2. Delsey, Tom. *The Logical Structure of the Anglo-American Cataloguing Rules: Parts I and II.* Accessible on the AACR web site at www.nlc-bnc.ca/jsc/aacrdel.htm (Part I) and at www.nlc-bnc.ca/jsc/aacrdel2.htm (Part II).
3. For an analysis of key issues relating to display of search results in the online catalogue see Allyson Carlyle, "Fulfilling the Second Objective in the Online Catalog: Schemes for Organizing Author and Work Records into Usable Displays," *Library Resources and Technical Services*, Vol. 41, No. 2 (April 1997), pp. 79-100.
4. For a description of one research project testing and evaluating a user interface supporting simultaneous access to a number of library catalogues through Z39.50 see F.H. Ayres, L.P.S. Nielsen, and M.J. Ridley, "BOPAC2: A New Concept in OPAC Design and Bibliographic Control," *Cataloging and Classification Quarterly*, Vol. 28, No. 2 (1999), pp. 17-44.
5. *The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery.* Release 1.1. June 2000. Accessible via the UKOLN site at http://www.ukoln.ac.uk/interop-focus/bath/1.1/

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Comments on Tom Delsey's Paper, "The Library Catalog in the Networked Environment" for the Library of Congress Bicentennial Symposium "Bibliographic Control in the New Millennium", Washington, DC, November 15, 2000

Jennifer Trant, Executive Director,
Art Museum Image Consortium and Partner, Archives & Museum Informatics

Final version

I was reticent about accepting an invitation to comment at a Library of Congress conference on the future of cataloging in the new millennium, for I am neither a library cataloguer nor an information scientist. My knowledge of bibliographic tradition parallels my knowledge as a Canadian, of the American electoral system: alternating glimmers of understanding and total bafflement.

But when I considered the topic, I began to think that I might have something to bring to the discussion. I am, for example, an experienced cataloguer (though the things that I learned to describe are works of art). I've been involved in consortial efforts to define the nature of a "catalogue description" (the Categories for the Description of Works of Art), and I now direct a consortium (AMICO - the Art Museum Image Consortium) that is building what might be considered a union catalog. Many of the issues that have been addressed in these discussions parallel those of the library community - but they've taken place in a parallel universe.

I'd like to highlight some of these issues within the context of the Tom Delsey's subtly expressed repositioning of the library catalog in the networked environment. He challenges us to rethink the nature of the catalog and what it describes. To fully understand the implications of this repositioning, we need to consider how the catalog is used, and when it is called upon in the research process.

The Nature of the Catalogue

First of all, I'd like to thank Tom for shifting the focus of the discussion from "bibliographic control" to the nature and purpose of the catalogue itself. The phrase "Bibliographic control" conjures up narrow connotations of the physical management of a distinct kind of object. The "networked catalog" shifts the our attention to a the role of description and access in the nexus between the collection and the user. But we have to be careful about the tempting inversion "cataloging the Web", for it may be the same kind of malapropism as "MARC cataloging". Just as librarians don't catalog with MARC, they don't describe the Web, but information resources that are available on the Web.

Conceiving of the catalog as an interface to networked information requires re-examining its content and its structure.

Many museums are now facing this question, as they consider the implications of putting 'collections online'. They are struggling to re-purpose collections management systems into public access systems. With both the catalog and the "object" delivered digitally, the boundaries between the catalog and the information resource it describes have become blurred. What was once considered a tool for managing a physical collection is becoming a means for presenting knowledge about that collection. With this change must come a parallel adjustment in the content and nature of the catalog itself. We are learning to wrap the individual descriptions of discrete objects with the context helps users make meaning. The data elements that enabled us to answer the questions 'What is it?" and "Where have I put it" don't fully satisfy the need to interpret the nature of a digital object, and help the researcher understand what a museum object means, and how it relates to other things.

What does the Catalog Describe?

Traditionally, the library catalog reflects a series of selection and acquisition decisions, based upon a collections development policy. Tom's paper hints that the networked library catalog may contain additional information. And he points us to the possibility that all of this information might not have to be supplied by cataloguers. Data from abstracting and indexing services may provide a level of granularity not achievable given the economics of traditional cataloging. Analytics virtually integrated into the catalog from other sources would allow the researcher to gain access to the unit of information appropriate for her task (for she'd like to find the article not the journal issue). Links to online texts enable the delivery of the resource itself.

Another challenge to the traditional conception of the bibliographic control, however, is that the resources likely to be used in a networked information space may or may not be "bibliographic". Formally published writings are now integrated with drawn or digitally photographed images, recorded sounds, reconstructed models, mathematical simulations in a fluid digital space and these new genres require a different methods and structures for their description.

These new genres raise new issues: much of the information required to adequately document such networked information resources is extrinsic to the resource itself. Even if a digital object could in some way technically 'self-describe' through a declaration of embedded metadata, much intellectual description becomes a matter of assertion. The catalog record begins to represent opinion, rather than fact. The catalog itself, becomes a publication -- its contents a compilation distinguished by their selection, arrangement, authority, authenticity and interpretation. This is certainly the direction the AMICO Library (http://www.amico.org/) is moving. Interestingly these characteristics are shared by the kinds of catalog that one often encounters in the art world: the Exhibition Catalog, the Permanent Collection Catalog and the Catalog Raisonnée all embody scholarship and opinion as much as they represent 'fact'. As document genres they sit on the boundary between metadata and data itself.

How is the catalog used?

It was helpful to be reminded in Tom's paper of the Functional Requirement s for Bibliographic Records: to find, identify, select and obtain. These processes seem to have their online equivalents in the realm of information discovery and retrieval. Much discussion of metadata in the Web environment has focused on this first step in the research process, finding relevant resources. (Cross-domain resource discovery was one of the motivators of the Dublin Core initiative.)

But information discovery isn't an end in itself. The act of obtaining information doesn't answer the researchers' question. Users of networked information resources have come to expect seamless support of their entire research process. The catalog and the information resource blur when the interface to both is the web browser. The digital library catalogue becomes the means for the delivery of dynamic digital content, requiring us to reassess the functional requirements of the library catalog. We're asking those catalog records to do much more than they used to! Tom hinted at this when he spoke about access management. Here, we could look to the archival community, who have long administered restrictions on access to collections, and to museums, whose relationships with contemporary artists offer another model.

The intersection between the catalog and the research process

Repositioning the catalog requires a model of when and where it is used. In a paper presented at a UK Office of Library and Information Networking Conference in 1998 David Bearman and I explored the inter-relationships between metadata requirements and the Humanities research process (paper online at http://www.archimuse.com/papers/ukoln98paper/index.html). We identified five broad phases in this process: Discovery, Retrieval, Collation, Analysis and Representation. Building on the ideas expressed in the Warwick Framework we wondered what kinds of metadata would be required at which point, in this iterative process, and how it might be supplied by a well designed library catalogue working in a networked information system.

Discovery and Retrieval are phases we are familiar with. Finding (Discovery) takes place in the public space of the networked library catalog. The researcher identifies resources of interest, finds the available copies that are nearest, or most convenient, or most suitable. During retrieval, this content is moved from the library space to the users space, to enable further use. Already, in supporting digital retrieval we may be adding requirements for technical elements to our metadata, those required by the user to judge which of many formats might be most appropriate.

Librarians have not concerned themselves with what users do, and often philosophically denied any knowledge of this area of activity. But we need to consider use if we are to support it. Use can be broken down into a series of individual functions with specific characteristics, each of which require or generate metadata: *collation* (integrating new resources into existing ones), analysis (deriving, creating or assigning meaning) and *re-presentation* (the publication or re-distribution of new knowledge. This process is cyclical; the act of re-presentation creates a new resource to be discovered. If metadata is managed throughout this process, then the description of the new resource is much easier.

Different kinds of metadata will have a role to play in each of these phases. Much of this is metadata that is likely to find its way into a the catalog of a distributed digital library. Further research and discussion about the nature of that catalog and the way that users interact with and use digital information resources and their descriptions is critical to the creation catalogs with utility. The value that the library profession could add to these new kinds of catalogs may not be in locally created catalog records -- other authors have drawn our attention to initiatives producing metadata along with networked information resources and the bar-code scanning cat distributed with Wired magazine is now touted as a way to catalog books (see http://www.wirednews.com/news/gizmos/0,1452,39139,00.html)

Seeing the role of the library catalog as that of a mediator and provider of access to networked information, rather than as a management tool for a repository of books requires a more active management of resources and relationships. Architectural solutions than enable the incremental integration of disparate and distributed resources are key to enabling us to "fast track" the building of catalogs of networked information (pouring concrete in the foundation before all the details are designed is not unusual in large architectural projects). What is key is up-front exploration of the knowledge structures of the disciplines we serve, and how do they intersect.'.

Throughout our discussions of the future of the library catalog we need to remember that users don't search the catalog in order to find a catalog record. They aren't even really looking for a book -- they are looking for information resources that help them answer their questions and accomplish their tasks. The challenge for the digital library catalog is to provide the right information about useful resources, at a time and in an environment that supports user processes. Looking outward is key to repositioning the library catalog within the networked information environment.

Library of Congress
January 02, 2001
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Priscilla Caplan

Assistant Director for Digital Library Services
Florida Center for Library Automation
2002 NW 13 St., Suite 320
Gainesville, FL 32609

International Metadata Initiatives: Lessons in Bibliographic Control

About the presenter:

Priscilla Caplan has been at the Florida Center for Library Automation since Aug. 1999. Previously, she served as Assistant Director for Library Systems, University of Chicago Library, from Aug. 1993-July 1999, and as Head, Systems Development Division, Office for Information Systems, Harvard University Library, from July 1985-July 1993. Her professional activities include being co-chair of the Dublin Core (DC) Standardization Working Group (1999-) and member of the DC Advisory Committee (1998-); chair, National Information Standards Organization (NISO), Standards Development Committee (1997-) and member of the NISO Board of Directors (1998-); Lecturer, Dominican University, School of Library and Information Science (July 1998-July 1999) ; Director, CUIP Digital Library, Chicago Public Schools/University of Chicago Internet Project (Nov. 1997-July 1999); member of the Digital Library Federation, Architecture Committee (1998-1999); and member (1991-1993, 1993-1995 terms) and chair (1995-1996) of MARBI. Caplan has written extensively on metadata and related issues which have been published in The Cybrarian's Manual 2, D-Lib Magazine, Public Access Computer Systems Review, The Serials Librarian, and Cataloging & Classification Quarterly

Full text of paper is available

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

The decade of the 1990s saw the development of a proliferation of metadata element sets for resource description. This paper looks at a subset of these metadata schemes in more detail: the TEI header, EAD, Dublin Core, and VRA Core. It looks at why they developed as they did, major points of difference from traditional (AACR2/MARC) library cataloging, and what advantages they offer to their user communities. It also discusses challenges to implementers of these schemes and possible future developments. It goes on to identify some commonalities among these cases, and to attempt to generalize from these some lessons for developers of metadata element sets. It concludes by suggesting we also look carefully at emerging schemes being developed by publishers in support of electronic commerce and rights management, and think seriously about the implications of commodity metadata upon our traditional bibliographic apparatus.

Robin Wendler, commentator

Office for Information Systems
Harvard University Library
Cambridge, MA 02138

About the commentator:

Robin Wendler is Metadata Analyst in the Harvard University Library Office for Information Systems (OIS). For the past two years she has worked on the Library Digital Initiative (LDI), a program to develop the infrastructure Harvard libraries need to acquire, manage, deliver, and preserve digital materials as systematically as other formats. She advises on the design of widely diverse kinds of metadata both to the LDI development team and to project managers developing digital content. Recent projects have focused on administrative metadata for digital audio, digital repository metadata, and visual resource cataloging.

From 1989-1998 she was the Bibliographic Analyst in OIS, providing functional analysis on the use of MARC formats in Harvard's local integrated library system, the import and export of cataloging data, and the specification of technical services functions. Prior to coming to Harvard in 1988, she was an original cataloger for art and architecture at the University of Maryland, College Park.

She currently sits on the RLIN Database Advisory Group and the CC:DA Task Force on the VRA Core Categories. Other professional activities have included MARBI (1995-1999), Co-chair of the ALCTS/LITA Institute "Managing Metadata for the Digital Library: Crosswalks or Chaos" (1998), Digital Library Federation Task Force on Metadata (chair, 1997), consulting on MARC formats for a library systems vendor and on multilingual thesauri for a European library consortium.

Full text of commentary is available

Library of Congress
July 5, 2000
Comments: lcweb@loc.gov

International Metadata Initiatives: Lessons in Bibliographic Control

Priscilla Caplan
Assistant Director for Digital Library Services
Florida Center for Library Automation
2002 NW 13 St., Suite 320
Gainesville, FL 32609

A blooming garden, traversed by crosswalks, atop a steep and rocky road

Final version

Library historians are likely to see the1990s as a decade of particular excitement, creativity and change. It will certainly be known for the rise of the World Wide Web, and as the decade that the Digital Library was invented. It may also be known for an almost explosive proliferation of metadata schemes. The first draft version of the Text Encoding Initiative (TEI) Guidelines, including the definition of the TEI header, was distributed in 1990. The first version of the FGDC Content Standard for Digital Geospatial Metadata was released in 1994. The workshop that drafted the original Dublin Core Metadata Element set was held in 1995. The alpha version of the Encoded Archival Description (EAD) was released in 1996. The Core Categories for Visual Resources version 2.0 was released by the Visual Resources Association in 1997. The Data Documentation Initiative was established in 1995 and released an XML version of the DDI metadata standard for social science data resources in 1997. The learning resources community produced both the Dublin Core-based Gateway to Educational Materials (GEM) element set in 1998 and the IMS Meta-data Specification in 1999. And so on; this list is only a sampling. In the metadata garden, truly a thousand flowers are blooming.

This has been a mixed blessing for libraries, presenting (as most innovations do) both opportunities and challenges. On the positive side, it has given us new options for describing materials that are poorly served by the AACR2/MARC suite of standards, and it has created a renewed sense of intellectual excitement in resource description. At the same time, these new formats have placed new burdens on the library profession. There are mature, well-developed tools for creating and managing traditional cataloging [1]. There is in fact an entire industry dedicated to its support -- the integrated library system is after all integrated around the multipurpose bibliographic database. Now suddenly we are confronted by content standards with no syntax and with data structures that we have no systems to support. Suddenly we are charged with supporting any number of schemes, not to mention maintaining registries of them and crosswalks between them. Suddenly there is an expectation we can control and give access to metadata created by organizations outside of our own library community.

This paper looks at a subset of these metadata schemes in more detail: the TEI header, EAD, DCMES, and VRA Core. It looks at why they developed as they did, major points of difference from traditional library cataloging, and what advantages they offer to their user communities. It also discusses challenges to implementers of these schemes and possible future developments. It goes on to identify some commonalties among these cases, and to attempt to generalize from these some lessons for developers of metadata element sets. It concludes by suggesting we also look carefully at emerging schemes being developed by publishers in support of electronic commerce and rights management, and think seriously about the implications of commodity metadata upon our traditional bibliographic apparatus.

The Text Encoding Initiative (TEI) Header

The TEI Header is a good place to begin because it is basically bibliographic, in a narrow sense of the word. Encoded texts are fundamentally like books in a way that art slides, museum objects and satellite data are not. Many texts marked up according to the TEI guidelines are based on printed books for which AACR2/MARC catalog records exist. The developers of the TEI specification were well aware of libraries and the principles of bibliographic description. Under these circumstances it is not unreasonable to ask why the TEI header was developed at all. Why didn't the Text Encoding Initiative rely on library catalog records, and put their energies towards modifying traditional cataloging to better accommodate TEI-encoded texts?

The answer, to a large extent, was a matter of workflow. The TEI drafters envisioned that the same individuals who marked up electronic texts would be creating metadata for them, and that these individuals would not be librarians but rather humanities scholars. These scholar/encoders might be experts in their own areas but they could not be expected to learn cataloging rules, so the TEI guidelines quite deliberately do not require any cataloging knowledge. On the other hand, the drafters equally deliberately designed the header to provide a trained cataloger the information he would need to create a good cataloging record. [2] The header areas are based on ISBD, but rules for obtaining and representing the content are not prescribed.

Similarly, scholar/encoders could be expected to know SGML markup, so it was natural to represent the metadata content in SGML rather than MARC. Using SGML in turn allowed the metadata to be embedded in the TEI document itself, processed by the same software, and searched within the same retrieval system. In theory, if a standalone record was required, the header could be used to algorithmically create a MARC record for importing into the library's catalog system.

In fact, events did not turn out exactly as envisioned. Most TEI texts are created not by scholar/encoders, but by the staff of projects or electronic text centers closely associated with libraries. In many cases header data is created or reviewed and revised by librarians. This has led to a perceived need on both sides to bring the header more in line with traditional cataloging. Librarians have found the "leg-up" provided by the header to be of limited usefulness. A study by the CC:DA Task Force on Metadata and the Cataloging Rules analyzed the TEI header as a source of cataloging data and concluded, not surprisingly, that the data is directly usable only to the extent that the encoder followed cataloging rules. [3] Automatically derived MARC records are problematic for this reason, and cannot be integrated into library catalogs without review.

At the same time, a desire to support searching across multiple collections, or even to share TEI data between institutions, has provided an impetus for more consistency in both content guidelines and encoding practices. The Oxford Text Archive sponsored a meeting in the fall of 1997 which acknowledged both the need for greater compatibility with traditional cataloging and greater consistency in practice between electronic text centers, resulting in the draft of a guide to good practice. The following year a workshop on TEI and XML in Digital Libraries held at the Library of Congress charged a working group to "recommend some best practices for TEI header content and review the relationship between the Text Encoding Initiative header and MARC", resulting in a draft TEI/MARC Best Practices document. [4] (Interestingly, most studies of the TEI header have focused on its compatibility with traditional cataloging and its usefulness in relation to a library catalog system. Whether the TEI conventions, designed to be useable by scholar/encoders, are more or less useful than traditional cataloging for scholar/users, has not to my knowledge been studied.)

The TEI Header is not, of course, directly analogous to a catalog record, and supports a number of data categories which cannot be mapped to MARC or can only be loosely mapped to note fields. The change history section (<revisionDesc>) provides a structured way to log changes made to an electronic text, including date, responsible party, and nature of change. The elements for describing the source on which a TEI text is based (the <sourceDesc> within the <fileDesc>) allows a detailed and richly content-designated description which goes far beyond the MARC 534, particularly for non-print sources such as the spoken word, audio or video recordings. The encoding description (<encodingDesc>) section provides a place for lengthy and detailed description of the encoding of the electronic file, including data about the project which created it, the purpose for which it was created, transcription practices followed, editorial decisions made, and SGML tagging applied.

The encoding description area of the header is notable because it supports a function not addressed by IFLA's Functional Requirements for Bibliographic Records (FRBR) [5]: the ability to make use of the resource. FRBR describes four generic user tasks that catalog records must support: to find materials that correspond to the user's stated search criteria, to identify an entity, to select an entity appropriate to user needs, and to acquire or obtain access to the entity. This "bibliographic" approach to metadata has been contrasted to the approach taken by computer scientists, which puts more emphasis on the management of data, including support of data use, data sharing, data security, and data integrity functions. [6] The TEI header contains elements of both traditions, treating the electronic text as both an object to be discovered and a data file to be used and managed over time.

In sum, the TEI header contains bibliographic information supporting resource discovery, and data management portions supporting use of the resource. Historically, the progression of the bibliographic portion of the TEI header has been toward greater consistency in encoding, greater compatibility with traditional library cataloging, and greater syntactical congruence with MARC. This makes sense in the context of an integrated information system, serving the user who may be interested in any and all versions of a work, including printed texts, electronic reproductions, and TEI encoded representations. At the University of Michigan, TEI headers are actually generated from MARC records, and in a Library of Congress implementation, bibliographic fields are left out of the header altogether. It may be that the TEI header will evolve to carry only minimal bibliographic description, with the bulk of the section being replaced by an external MARC record. The MARC record could then point to a TEI header containing detailed encoding, profile and revision information, much as collection-level AMC records are used to point to more detailed EAD finding aids today.

The Encoded Archival Description (EAD)

The developers of the EAD had both MARC and the TEI header available to them as models. Unlike the TEI header, however, the EAD was designed as an electronic finding aid to resources that would not necessarily be available in electronic form. While the EAD can be used to describe web-accessible collections, its primary purpose is to improve awareness of archival holdings in all formats.

The archival community had been using the MARC AMC format for some time to give high level access to archives and manuscript collections. Archivists found, however, that AACR2 was inadequate for archival description, and adopted instead Steven Hensen's *Archives, Personal Papers and Manuscripts* (APPM) for content rules. The principles of bibliographic description apply even less to finding aids, as Daniel Pitti has pointed out. [7] Bibliographic description represents a published item; archival description represents a fonds, or organically generated collection. Bibliographic description emphasizes physical characteristics; archival description emphasizes intellectual structure and content. Bibliographic description supports finding, identification, selection and access; archival description is evidentiary and must document provenance and original order. A distinction of practical importance is that bibliographic description is typically brief, stylized and flat. Archival description is typically lengthy, narrative, and deeply hierarchical, making SGML, and later XML, a more suitable transport syntax than MARC.

Although archivists generally follow principles for archival description in their local finding aids, and although the General International Standard Archival Description (ISAD(G)), a set of general rules for archival description, was adopted by the International Council of Archives in 1994, there is no ruleset equivalent to APPM specifically for finding aids. In the absence of an existing content standard, the developers of FindAid, the predecessor from which the EAD evolved, solicited examples of paper finding aids from the community. As repositories tended to contribute only the samples they considered their best, a de facto corpus of best practices was acquired and used as the basis for developing the FindAid DTD. While the EAD was designed to accommodate the range of practice that was found, it was also developed in the hope of encouraging common practices regarding data content.

Paralleling the TEI standard which has a header preceding the encoded text, the EAD is divided into two parts, a bibliographic header (<eadHeader>) and the marked up finding aid itself (<findAid>). [8] The finding aid describes the collection and the header describes the finding aid, reminding us that one man's metadata is another man's data. The header in turn has sections for describing the original finding aid (for example, its author and title), describing the encoded version (for example, whether it was created by OCR, retyping, etc.), and recording a revision history.

The EAD has been rapidly, widely and internationally embraced, particularly by university archives and special collections departments within academic libraries. Certainly one source of this success is the ability of the EAD to accommodate existing archival practice, rather than forcing practice to conform to the constraints of a data format or syntax. [9] Because of this congruence, it has been possible to convert paper finding aids with some success, and something of a cottage industry has arisen in providing vended conversion services. The EAD also appears to be filling a void in tools for detailed collection description, as institutions are applying it to collections of all sorts, not only those controlled archivally.

Adoption of the EAD has been notably slower outside of academic institutions. State archives, for example, still rely almost exclusively on collection-level MARC records. A meeting of the Southeastern Archives and Records Conference in 1999 concluded that the "EAD is not useful unless there is substantive information at lower levels, and most state records series have only box inventories, with the frequent exception of governor's records." [10] The SGML-based structure also requires specialized editing tools and software for search and display that can present a barrier to implementation at smaller institutions. While the scholar/encoders of the TEI might be expected to have a research interest in SGML, most archivists would have no reason to be familiar with this encoding apart from the EAD. RLG and the Society of American Archivists (SAA) have been proactive in sponsoring intensive training, which is almost a prerequisite to implementation.

A major strength of the EAD -- its ability to represent complex finding aids with a high degree of content designation, while accommodating a wide range of local practices -- can also be a drawback. Although a relatively small number of tags are required, the tagset itself is extensive, and every repository must arrive at its own set of guidelines for which tags to use, how they may be used, and how data may be represented within them. Because the EAD does not include and is not directly correlated with established content rules, this allows for some creativity, and there is wide variation in practice. Widespread implementation of the EAD has been followed almost immediately by the desire for union access. In 1998 RLG launched its Archival Resources service, a union catalog of distributed collection guides. Using a registry of contributors and a customized harvester, the service collects and indexes EAD and non-EAD finding aids. In 1998 and 1999, the Digital Library Federation undertook a project to develop a Distributed Finding Aid Search System (DFAS). DFAS implemented Z39.50 search and retrieval across distributed EAD repositories as an alternative to the union catalog approach. Both Archival Resources and DFAS found the diversity in encoding practice to be a major problem. A report of the DFAS project concluded, "Our research has highlighted the problems caused by the lack of standardization in the application of EAD to finding aids, yet that lack of standardization is not easily overcome given the diversity of the underlying documents." [11]

The EAD has very clearly encouraged archivists to conceptually reexamine the logic, structure and content of their finding aids. In some cases has inspired repositories to reengineer their finding aids for more effective web-based use. It appears that the next phase in EAD development will be the establishment of common guidelines, including Z39.50 profiles and Best Practices for encoding particular types of finding aids. Changes in the EAD DTD itself may be necessary as use expands beyond the academic community that invented it, and as more experience is gained in representing collections of both digital and non-digital content.

Dublin Core Metadata Element Set (DCMES)

The DCMES is unusual among metadata element sets in the generality of its application and use. In contrast to other schemes which target particular types of materials and particular user communities, DCMES can be, and probably has been, used to describe nearly any type of information resource.

Like the TEI header and the EAD, the DCMES has evolved in unexpected ways. Though originally envisioned as a mechanism for encouraging authors to supply metadata for their own publications, the vast majority of use is from projects associated with libraries, cultural heritage institutions and government agencies. Originally intended to support description and discovery of what Clifford Lynch has called the "dark matter", or largely invisible content, of the Web, DCMES has found a multiplicity of other applications. It has been particularly useful in support of interoperability -- retrieval across multiple existing metadata stores. In this capacity DCMES has been used as a minimal set of commonly understood access points for cross-domain searching, as a common extract format for creating union catalogs, and as a searchable entry point to local files of more complex metadata. An emerging use in Open Archives and related initiatives is as the basis of an extract format for harvesting metadata from dissimilar repositories.

The most astonishing thing about DCMES is the pervasiveness of its adoption. Although the Dublin Core website maintains a list of DCMES-based projects, this barely hints at the number of implementations worldwide using or somehow based on the Core. One reason this is possible is because DCMES allows, even encourages, the use of local extensions. The basic model is to use DCMES elements where they apply, and supplement them with domain- or application-specific elements where needed. XML namespaces, which are supported in RDF and in the emerging specification for XML Schema, provide a practical means for implementing this type of combination.

Ironically, this same use highlights a weakness of the DCMES as a building block for other metadata schemes. Much has been made of the "Lego"(tm) model, in which Dublin Core elements can be snapped into other schema as appropriate. However, Legos (tm) require an extreme degree of precision, and for this approach to work, DCMES elements should be related to a data model that can be precisely described. The DCMES, however, developed organically, and attempts to apply a more rigorous data model after the fact have had to contend with inconsistencies already present in the element set. This has led to some tension between the need to maintain stability for existing implementers on the one hand, and the desire to move the element set towards greater logical consistency on the other.

Practically, most projects using DCMES have found its lack of specificity to be a problem. DCMES 1.1 gives only the broadest description of semantic categories; there are no rules for how to determine or represent content, and only the most general guidelines are available in draft status from the website. As a consequence, projects using DCMES for resource description find it necessary to develop their own conventions, a difficult and time-consuming endeavor. A working group charged with developing a user guide found that although librarians tended to be frustrated by the lack of a ruleset, other sectors were not, and there was no general consensus that common content rules were either necessary or desirable. In any case, the huge diversity of applications argues against canonical guidelines. The expectation is that communities sharing particular resource needs will get together to develop domain-specific rules. However, this is itself an arduous process. The CIMI Guide to Best Practice, for example, now available in version 1.1, took three years, a testbed implementation and extensive community review to accomplish. [12]

Most projects have also found a need for some refinement of the very general DCMES semantics, ordinarily referred to as "qualification". An initial set of qualifiers formally approved by the Dublin Core Metadata Initiative (DCMI) is in final draft status at the time of this writing. (Exactly what this means in terms of compliance for applications is still somewhat unclear.) Applications and communities are encouraged to develop their own qualifiers, as with extensions, and to submit these to the DCMI for review and approval. However, much of the apparatus required to support consistent and confident use of qualifiers is still outstanding, including clear guidelines for constructing valid qualifiers; a registry identifying approved, not-yet-approved-but-valid, and invalid-but-needed-by-some-community qualifiers; and a mechanism for approving new qualifiers.

The DCMES began as a grass-roots movement, independent of existing organizations and without a formal structure to manage it. Over time the DCMI has evolved alongside the DCMES to be "responsible for the development, standardization and promotion of the Dublin Core metadata element set." [13] However, the very diversity of Dublin Core implementers makes it extremely difficult to achieve consensus on any but the most basic issues. Also, apart from a skeletal directorate, participation is largely a volunteer effort. Unlike other standards discussed here, the DCMES is not a program of any larger organization that sees maintenance of DCMES as part of its core mission. And, although DCMI procedures are modeled after the W3C, the DCMI, unlike W3C, has no formal membership with the corporate commitment and financial support that entails. The most critical factor in the future of DCMES is whether a working organization can be achieved to manage the change process and to produce the documentation, support structures, and policies required by an international community of implementers holding very little in common.

Visual Resources Association (VRA) Core Categories for Visual Resources

Like the DCMES, the VRA Core was conceived as a core set of elements that particular applications could enhance with additional elements as needed. In contrast to the very comprehensive Categories for the Description of Works of Art (CDWA), the VRA Core was designed as a moderate set of elements which, if commonly supplied, would support the sharing of data for visual materials. Historically, catalogs or databases of visual materials tended to be institution-specific, using locally-defined data elements, formats and authorities. There was a great redundancy of effort, as every institution cataloged their own collections of slides based upon the same works of art. Widespread Internet use brought increasing pressure to share data, not only to help users find materials but also to create an environment in which works could be cataloged only once. As one of the drafters of the VRA Core described it, "The point of the exercise ... was to develop a set of elements that all visual resources curators could use to share information about the works of art so that they would not have to repeat the research process for each work represented in his/her collection." [14]

The VRA Core is still evolving fairly rapidly, moving towards a more generalized and more flexible model of the visual materials universe with each version. Version 1.1, which was never widely implemented, was at heart based on the collection of art slides, the basic model being an art object that was not held by the cataloging collection and a slide of the art object that was. The original Core consisted of descriptive elements, or "categories", to describe the object, the creator of the object, and the surrogate. Version 2.0 was a deliberate attempt to generalize the element set to accommodate non-art objects and to give greater weight to the surrogate, the term itself generalized to "visual document". VRA Core 2.0 defined 19 "Work description categories" and another nine "Visual Document description categories". Both version 1.1 and 2.0 attempted to accommodate the practical experience of catalogers of visual materials, that in describing a slide or photograph in their collections, they were simultaneously describing the original work in some other medium. However, this pairing breaks down very quickly into far more complex relationships: not only can there be multiple representations of the same work (a slide, a photo, a digital image), but some of these may be surrogates of others (the digital image is made from the photo), while some may be works in their own right (the photo was taken by a well-known artist). Works may exist as parts of wholes (a stained glass window in a building) or as parts of collections; visual documents may exist in collections, may encompass multiple works (a photo of two buildings), and so on.

Version 3.0 acknowledges this complexity by abandoning the attempt to separately describe works and various representations of them; it simply defines 17 categories that can be used as appropriate. Although it retains the conceptual distinction between a work and representations of the work (now called "images"), it embraces the "1:1 principle" popularized by Dublin Core, that a single set of metadata elements should describe a single entity, and it assumes that records describing images will be linked to the related works. It also incorporates the Dublin Core concepts of elements and element refinements, or "qualifiers"; for example, what in version 2.0 were independent categories for Creator and Role are in version 3.0 one category Creator with the qualifier Role.

Version 3.0 is too recent for widespread implementation, although many of the concepts it represents were previewed in Harvard's Visual Image Access (VIA) application. However, in general the VRA Core has been gratefully, even hungrily, received by visual resources curators. One attraction has certainly been its latitude in addressing both an original work and a derivative surrogate, something very difficult to accomplish in traditional library cataloging. Other attractions have included its focus on visual materials with distinct categories for concepts such as measurements, material and technique, and its flexibility in accommodating local cataloging practices. It has been found to be applicable to works of architecture, non-art images, and other domains beyond the art history slide collection. A paper by Marcia Lei Zeng describes how the VRA Core 2.0 was chosen over MARC and Dublin Core for cataloging a museum collection of historical costumes. [15]

On the other hand, while the VRA Core has been used for describing materials in institutional collections with some success, the visual resources community has some way to go toward achieving the goal of sharing information about works. The lack of standard cataloging rules and common authority schemes for content presents a major barrier to interoperability. Not only is the historical insularity of visual resources collections reflected in any number of purely local vocabularies and classification schemes, but, as visual materials cover a wide range of territory from pottery shards to buildings, a large number of specialty thesauri are in use. Inconsistency in the use of authorities was noted as the major problem in VISION, a testbed for the VRA Core version 2 developed by the VRA, the Getty Information Institute, and the Research Libraries Group. The testbed also revealed a main concern of participants was mapping to and from local databases, a sign that, for early implementers at least, local structures were still their foremost concern. The rapid evolution in the structure of the VRA Core indicates the community has yet to develop an underlying data model to support the complex relationships these materials exhibit. The future of the VRA Core probably depends less upon further improvements in the Categories themselves than on whether their existence serves as a catalyst for the development of a shared data model and content rules.

Functional Requirements for Metadata Records

It could be argued that beyond being intended for electronic description of information resources, the metadata schemes discussed above have little in common. They have different intended users and different intended uses. They are to varying degrees "bibliographic", in terms of being designed to support the Functional Requirements for Bibliographic Records. Some are defined in terms of DTDs, while others are semantic categories independent of syntax. In fact, none of these schemes are exclusively (and some not even primarily) intended for controlling electronic resources: to varying extents they describe paper and artifactual resources as well as digital. One should therefore be cautious about making inferences regarding lessons for bibliographic control of the web. Nonetheless a few generalizations are cautiously offered.

For starters, in no case did the actual creators, users and uses of these schemes turn out to be just what their developers anticipated. Metadata takes on a life of its own. Metadata schemes need to be seen as organic creations evolving in response to a changing environment, with the implication that a mechanism for effecting and controlling this evolution needs to exist. Ideally, such a mechanism is perceived as legitimate and authoritative, has a well-defined structure and process, gathers broad input from affected communities, and controls the rate of incremental change to ensure it is neither too fast for implementers to accommodate nor so slow as to present a barrier to effective use. It is arguable whether any of these emerging metadata schemes have managed to put such a mechanism into place, and it will be interesting to see whether and how these develop. In fact, it will be interesting to see whether the mechanisms governing change to traditional cataloging that have proved sufficient for a universe of print and other fixed, physically distributed media are in fact sufficient to accommodate the control of electronic resources in the rapidly changing network environment.

Another feature shared by all of these schemes is that none of them include or are based on rules for determining and representing content. What we have learned from this is that metadata schemes without content rules are not very useable. Implementers are forced to expend significant time and effort developing their own local guidelines to ensure some consistency in content and encoding within their own resource description projects. As usage becomes more widespread the desire arises to share metadata or to implement union search over a number of repositories, at which point the plethora of local guidelines immediately becomes a hurdle to overcome. In the next phase of maturity, implementers struggle communally to work out common use profiles to guarantee some minimal level of interoperability. Implementers' agreements in turn raise problems of their own: how are they publicized, who officially "owns" them, how are they maintained over time, how to accommodate (or prevent) the development of multiple competing profiles?

In the case of TEI, we see a movement towards greater conformance with traditional library cataloging, while the EAD is serving as an impetus for the development of content rules for finding aids, and there is some hope that the VRA Core will do the same for visual resources. The approach of the DCMES, with its many diverse user groups, has been to encourage the development of community-specific (as opposed to implementation-specific) guidelines and to encourage the use of existing authorities designated with a "scheme" qualifier. In all of these cases, but perhaps most intriguingly with DCMES, what we have been seeing, if we've been paying attention, is the re-invention of cataloging. For example, an extensive exchange concerning the nature of the distinction between Creator and Contributor took place on the main Dublin Core discussion list in the spring of 1999; it explored with some nuance the need to capture primary intellectual responsibility. On the negative side, we can see these communities slowly, painfully and with many false starts rediscover principles that librarians have understood all along. On the positive side, it will be constructive to learn from what they retain and what they throw away, because they are directly confronting what is necessary and feasible to meet the needs of users in the Internet environment.

One of the areas where guidelines are most needed is in how to handle works that are known to be available in different file formats (e.g. LaTeX and PDF) or different manifestations (e.g. a photograph and a digital image made from it). The IFLA model of work, expression, manifestation, and item is useful in untangling multiple versions conceptually, as is the principle of "1:1" espoused by DCMES and the VRA Core. However, it is by no means clear how to apply these practically, as discussions in both communities makes evident, or what mechanisms in supporting systems or in metadata schemes themselves might be required to effect reasonable retrieval and display in this context. [16]

Functions of metadata beyond resource discovery and identification appear to be especially important for electronic resources. The question here is which of these functions are best supported in descriptive schema and which in separate, complementary schemes. Restrictions on access and use, for example, can be seen not so much as a property of a resource but of the intersection of resource, user and use; most emergent descriptive schemes have shied away from extensive recording of rights and permissions, leaving these to other systems. Similarly, there has been quite a bit of activity in defining element sets for administrative and technical metadata useful in managing and preserving digital data over time, including sets defined by the RLG PRESERV, CEDARS, NEDLIB and CURL initiatives.

On the other hand, metadata schemes focused on complex electronic resources tend to include information needed to actually use the resource. The TEI header, for example, allows for lengthy description of encoding practices, and the Data Documentation Initiative (DDI) DTD for describing social science datasets, contains a "data files description" section for a detailed description of the format, size and structure of the datafiles. Metadata documenting the creation and maintenance of the metadata itself also appears to be an important and legitimate need, especially as SGML/XML-based formats encourage lengthier descriptions, maintained over time. Mechanisms for ensuring the authenticity of metadata will almost certainly be required.

Another apparent point of commonality seems to be an inclination to move information about agents (human and legal) into separate files, defined by separate metadata element sets. In the archival sphere, work is proceeding on an SGML encoding of archival authority records based on the International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR(CPF)). In contrast to LC name authority records, which primarily identify the authorized form of name and contain little other information, these records "describe fully the attributes of the creator needed to appreciate the context of creation of a body of archival documents." [17] By developing an SGML encoding for ISAAR(CPF), archivists will be able to divorce the capture and maintenance of contextual information from the description of the archival entity in the EAD itself. The VRA Core has also evolved away from carrying detailed information about the creator. Version 1.1. included a set of categories pertaining to the creator, including Creator, Nationality and Culture. In version 2.0, Nationality and Culture were redefined to apply to the work, with the recommendation that these data as applied to creators should be recorded in an auxiliary authority file. The DCMI is considering development of an "Agent Core", a structured set of metadata elements such as affiliation and address which properly pertain to the agent (Creator, Contributor, Publisher) as opposed to the resource. This proposal is congruent with an RDF data model where the value of the property "Creator", for example, is itself a resource with properties of its own.

If anything is clear from this it is that the metadata environment is becoming increasingly complicated for both the information provider and the information seeker. Not only are there more metadata schemes for different types of resources, but these schemes rely upon both implementers' agreements to restrict practice and upon local extensions to broaden it. In addition, metadata records created at different times by different agencies and located in different places may have to be integrated at various points of use. In traditional bibliographic environments, the primary form of coordination required has been between descriptive bibliographic records and authority files for names and subjects. Both bibliographic and authority files tend to be under a library's direct control, and headings are either stored redundantly or there are direct links between the two types of records. Despite the relative simplicity of this model, a huge amount of effort goes into its maintenance, and library systems seem rarely to do exactly what one would want. It remains to be seen how record creation, maintenance, retrieval and use will perform in this far more complex environment.

And now for something completely different...

To this point most of our efforts have been related to metadata schemes that have been developed by librarians, archivists, curators and other information professionals, or by government agencies or research initiatives such as the Data Documentation Initiative, the Federal Geographic Data Service, and the National Biological Infrastructure Initiative. To date we have not focused much attention on schemes coming out of the publishing community. However, these may ultimately have the greatest impact on traditional bibliographic description.

Several international efforts have been proceeding more or less simultaneously. The best-known in the library community is probably the INDECS (Interoperability of Data in E-Commerce Systems) project. INDECS was funded by the European Commission and supported by major trade associations representing record companies, music publishers, film companies, and book and journal publishers. The goal of the project was to create a framework for electronic trading of intellectual property rights in all media, and the primary product was a metadata model which is due for release in final form this summer. (Although the original project officially ended in March 2000, its work may be carried on by a not-for-profit membership organization.)

The INDECS model is essentially a semantic model for describing intellectual property, the parties that create and trade it, and the agreements that they make about it. The assumption is that many different metadata schemes will be developed and used by specific industries (for example, music and book publishers), and that it must be possible for this metadata to be exchanged between industries and reused in different contexts for global electronic commerce to thrive. INDECS attempts to distill the potentially infinite range of descriptive elements pertaining to rights into a defined set of generic, universally applicable attributes and values. Data can be exchanged between domain-specific metadata schemes if they follow or can be mapped to the INDECS data dictionary. The example often given is that the different schemes may recognize screenplay adapters, translators or musical arrangers, but translated to

INDECS, these are all specific examples of a generic category (contributor agent role) and value (modifier).

INDECS principles impose other constraints on metadata schemes as well. Because rights can be traded at any level of the IFLA model (works, expressions, manifestations, items) good descriptive metadata will not conflate these levels, and will provide for extensive, explicit linking between them. Because virtually any element of descriptive metadata can be an element of a rights agreement (except titles), the values of elements must be strictly authority-controlled and stored as unique, coded values. Because rights agreements depend on metadata, the authority for any item of metadata must be securely identified.

While work was proceeding on the INDECS framework, the Association of American Publishers (AAP) developed over a short period in 1999 a metadata element set for exchanging product information for the book trade. Called the Guidelines for Online Information Exchange (ONIX), the specification was released in version 1.0 in January, 2000. ONIX was a direct response to the enormous growth in online booksales, which has resulted in a need for publishers, booksellers and distributors to create and exchange vastly expanded metadata for saleable items. The introduction to ONIX 1.0 points out that "Books with cover images, descriptions, reviews and additional information online outsell books without that information eight to one."

The same month that ONIX 1.0 was published, the EPICS Data Dictionary version 3.02 was released by EDItEUR, an international book and serials industry standards group. EPICS was developed as a joint project of EDItEUR, the Book Industry Communication (BIC) in the UK, and the Book and Serials Industry Communication (BASIC) in the US. Like ONIX, EPICS is a metadata specification designed for exchanging product information, motivated partly by the rise of Internet bookselling, and covering bibliographic, promotional and trade information.

Work immediately began to unite the two efforts. A new version of ONIX, consistent with EPICS and intended for both U.S. and European implementation, was released in May 2000 under the name ONIX International 1.01.[18] EPICS has been redefined as a more comprehensive data dictionary of which ONIX is a book industry subset, and the broader EPICS is being expanded to other areas, starting with audio-visual materials. Both schemes will be maintained by EDItEUR under the direction of a single international steering committee. These are extremely fast-moving standards and it is likely there will be additional developments between the time of this writing and the Bicentennial Conference on Bibliographic Control. The comments below are based on EPICS 3.02.

The EPICS data dictionary was developed coterminously with the INDECS project and has increasingly adopted the INDECS data model; it is seen as one of the first INDECS-compliant metadata standards. It should be possible to map EPICS elements to generic INDECS elements. Also in keeping with the INDECS model, EPICS requires precise and granular identification of all data elements in the scheme, supports both text and authority-controlled codes for nearly all data values, and allows extensive relationships between the described object and other objects to be specifically recorded.

It is interesting to compare the semantics of the more bibliographic elements of EPICS with those of traditional library cataloging. For a title, for example, it is possible to identify the type of title (title on piece, constructed title, alternative title, ISSN key title, etc.) and various subelements within a title, such as "title prefix" (which would be noted as non-filing characters in a 245), title, subtitle, etc. The semantic overlap is imperfect but substantial. However, there are major differences in the conceptual structure. An EPICS title cannot carry a statement of responsibility such as carried in the 245 subfield c, which is in fact information strictly pertaining to contributors and their roles. Similarly a former title, classified as a title in MARC (grouped in the 24x block), would be classified in EPICS as an element pertaining to a related object.

There is also a difference in the approach to content rules. The AACR2 approach is to take care in the selection of recorded data. There are extensive rules governing the selection of main entry, the justification of added entries, the chief source of information for title, etc. The publishers' approach is to allow the recording of any data so long as the nature of the data is explicitly recorded. The names and roles of all contributors can be recorded, as can the names and types of all titles. Selection of the most appropriate contributor or title for a particular purpose is not a function of the creator of the metadata, but rather of the user of the metadata (most likely a computer program). On the other hand, the publishers lean more strongly towards the use of coded values from named authority lists for the representation of content.

One reason that these approaches to resource description differ is because the underlying functional needs for the metadata differ. The publishing community is far more concerned with marketing and with managing intellectual property rights, while the library community has a need to manage huge inventories over a very long period of time. Nonetheless, both communities need to support end-user discovery and identification of information resources, and there is great overlap in the user tasks that must be supported by basic bibliographic data.

Much of our attention to date has been focused on what we might call specialty metadata schemes. While this has helped to increase our sophistication and understanding of metadata issues in general, and while it has surely enhanced access to important categories of materials, I would suggest that it is time to look through the other end of the telescope and begin thinking about basic bibliographic metadata as a commodity, produced and exchanged by a number of communities in order to serve a number of purposes. We are already in an environment where readers are as familiar with amazon.com as with their library catalogs. We are already in an environment where libraries purchase catalog records from any number of sources, from OCLC's PromptCat to our approval plan vendors. We will soon be in an environment where most metadata is exchanged in XML: the publishers have already adopted it, and library systems are moving in that direction. In this context it makes very little sense to think that libraries, publishers, booksellers, distributors and vendors will all be creating incompatible, non-reusable bibliographic metadata.

I am not sure myself what it means to think about commodity metadata. Perhaps that is something that can be explored in the context of the Bicentennial Conference. However I do urge librarians to take a serious and objective look at the metadata schema emerging in the publishing community with the long-term goal of maximizing the interchangeability of data. It will not be enough to simply develop mappings between these schemes and MARC. Experience with the TEI header and with crosswalks from DCMES to MARC has shown that simply mapping from a semantic or syntactical element in one scheme to a comparable element in another does not guarantee the usability of the converted metadata.

I suggest that we work proactively with publishers to establish enough commonality between our respective rulesets to allow meaningful exchange and reuse of metadata. Can we establish common authority lists? (For example, libraries use MARC relator codes and ONIX International uses contributor role codes -- is there a compelling reason for these to be different?) Are there content rules which, if shared, would substantially benefit both communities? Are there content rules in traditional library cataloging that don't make enough of a functional difference to insist upon? I also suggest that we evaluate the additional metadata elements designed to support the book trade for their potential use in our own systems. Does content such as author biographies, book review excerpts, and dust jacket summaries provide useful access points for retrieval, or help a user select an item appropriate to his needs (both end user tasks to be supported in FRBR)?

In sum, I suggest that the key question as we enter the new millenium is not bibliographic control of Web resources, but rather bibliographic control of both digital and non-digital resources in the Web environment. I expect that the Web environment will be characterized by the development of competing search engines and retrieval models, a proliferation of commercial and non-commercial bibliographic services, and the dominance of XML as a transport syntax for both data and metadata. Evidence indicates that successful metadata schemes must be flexible enough to accomodate unexpected users and uses, must have responsive mechanisms for change, must be based upon or work in conjunction with shared content rules, and must allow clear relationships to be established between different works and manifestations. While refining specialty metadata schemes, we should also work towards the development of a system of commodity metadata that will enable economic exchange, reuse and repurposing of metadata for current trade publications in all media.

ACKNOWLEDGEMENTS

Many thanks to John Price-Wilkin and David Martin for their reading of and helpful comments on sections of this paper.

NOTES

1. This paper will use the term "traditional cataloging" to refer to resource description based on a suite of rules including ISBD, AACR2, LC rule interpretations, LC name authority, and the

MARC formats for bibliographic data. I do this not to reflect a value judgement for or against traditional cataloging, but only because some short-hand term is needed.

2. "It is the intention of the developers, however, to ensure that the information required for a catalogue record be retrievable from the TEI file header, and moreover that the mapping from one to the other be as simple and straightforward as possible." C.M. Sperberg-McQueen and Lou Burnard, eds., Guidelines for Electronic Text Encoding and Interchange (TEI P3) (Chicago; Oxford : Text Encoding Initiative, c1994.) p.137. http://www.uic.edu/orgs/tei/p3/.
3. Committee on Cataloging: Description and Access, Task Force on Metadata and the Cataloging Rules. Final Report. August 21, 1998. http://www.ala.org/alcts/organization/ccs/ccda/tf-tei2.html.
4. TEI/MARC "Best Practices", November 25, 1998 Draft. http://www.lib.umich.edu/libhome/ocu/teiguide.html.
5. International Federation of Library Associations and Institutions. Functional Requirements of Bibliographic Records: Final Report. September 1997. http://www.ifla.org/VII/s13/frbr/frbr1.htm#1
6. Kathleen Burnett, Kwong Bor Ng and Soyeon Park. "A comparison of the two traditions of metadata development." Journal of the American Society for Information Science 50(13):1209-1217, 1999.
7. Daniel V. Pitti. "Encoded Archival Description: An Introduction and Overview." D-Lib Magazine, 5(11) November 1999. http://www.dlib.org/dlib/november99/11pitti.html
8. Actually there are three sections; an optional section can be included to supply a more "publisher-friendly" title page than the header provides.
9. You can almost hear the surprised delight in early testimonials to the EAD, such as this quote from a talk by Susan von Salis, Schlesinger Library at Radcliffe College, to the RLG Forum in Toronto, 1997. "As I mentioned, most finding aids include common components such as provenance, scope and contents, and access restrictions. So..... the DTD includes these 'parts' as its elements! Markup itself is simply a matter of wrapping the correct tags around the proper text." http://www.lib.umb.edu/newengarch/InternetResources/vonsalisrlg/index.html
10. Report of the Emerging Descriptive Standards Group, Southeastern Archives and Records Conference, Columbia SC, May 23-25, 1999. http://www.state.sc.us/scdah/sarc41999.htm
11. MacKenzie Smith. "DFAS: The Distributed Finding Aid Search System." D-Lib Magazine, 6(1) January 2000. http://www.dlib.org/dlib/january00/01smith.html
12. Consortium for the Computer Interchange of Museum Information. Guide to Best Practice: Dublin Core, version 1.1, April 2000. Available from http://www.cimi.org/standards/index.html#FIVE.
13. Renato Iannella and Rachel Heery. Dublin Core Metadata Initiative - Structure and Operation. April 1999. http://purl.org/dc/about/DCMIStructure-19990531.htm
14. Lynda S. White. "Creating the VRA Core: The Critical Issues." VRA Bulletin, v.25, no.4 (Winter 1998), p. 34-40.
15. Marcia Lei Zeng. "Metadata Elements for Object Description and Representation: A Case Report from a Digitized Historical Fashion Collection Project". Journal of the American Society for Information Science 50(13):1193-1208, 1999.
16. Bernhard Eversburg summarized the principle of 1:1 with the following verse:

Make metadata one to one,
just one per item, is the task.
Rather less,
more's a mess!
"But what's an item", now you ask?
If that's in doubt, do none.
http://www.mailbase.ac.uk/lists/dc-general/1999-04/0117.html Nonetheless, after extensive debate over whether Ansel Adams or the scanning technician is the Creator of a digitized Adam's photo, the answer appears to be that the Creator is in the eyes of the beholder.

17. International Council on Archives. ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families. Ottawa : The Secretariat of the ICA Ad Hoc Commission on Descriptive Standards, 1996. http://dobc.unipv.it/obc/add/infap/archdes/isaar_e.html
18. ONIX International Version 1.01. http://www.editeur.org/onixfiles.html

---

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Musings on Priscilla Caplan's "International Metadata Initiatives: Lessons in Bibliographic Control"

Robin Wendler, commentator

Final version

Any attempt to review metadata initiatives on an international scale is daunting not only because of the sheer number of efforts but also because of the range of forms they take and the range of purposes they serve. Some metadata initiatives produce data element lists and data dictionaries, of course, but others produce encoding syntaxes, registries, controlled vocabularies and thesauri, rules for selecting or formulating the content of data elements, and so on. Metadata elements or values with the same semantics can exist in diverse schemes designed to support intellectual access, rights management, preservation, commerce, and structural representation. Complicating the picture is the fact that many metadata schemes encompass more than one of these categories to some extent.

Priscilla Caplan's paper focuses primarily but not exclusively on schemes which support intellectual access. She provides a very focused overview and an insightful analysis of the key metadata initiatives of relevance to libraries and to the kinds of organizations most closely allied with them, such as archives, museums, and publishers. Rather than comment on the initiatives individually, I'd like to highlight some of the conclusions she draws. These are essential points. Not only are they the logical lessons to be drawn from a review of these metadata efforts, but they reinforce the lessons libraries themselves have learned about creating and sharing metadata on a large scale.

"Approaches to resource description differ because the underlying functional needs for the metadata differ."

Access to information resources does not occur in some abstract space. Resources are described within a context, and a description of a resource reflects the perspective of that context. With the exception of the Dublin Core, each of the schemes Cilla analyzes has emerged from a specific community with a specific worldview and was developed to fulfill the requirements of that community. Those requirements often differ markedly from those that AACR2 and MARC were designed to support.

Cilla describes the marketing and rights management needs of the publishing community, and contrasts them with the need of libraries to manage huge inventories over long periods of time. Another example of how the same materials are described in very different ways comes from the image world. A given photograph held in a visual resource collection or an archive would receive radically different treatments in those environments than it would in a library catalog. The visual resource approach gives primacy to the subject of the photograph, as opposed to the photograph as an object per se, and requires a richer and more integrated way of managing relationships among various "works" and images of those works than MARC provides. Archivists, like libraries, tend to describe the photograph as an artifact. However, while they generally provide little information about an individual photograph, they do require the ability to express how it fits within the intellectual organization of a complex body of materials.

Seeing how differently another community conceives its information reminds us to be sensitive to their functional needs. In some sense, the contrast does, or should, make us see our own in a new light. This means both questioning our own assumptions about how resources should be described and why, but also valuing the requirements which survive such scrutiny. When we evaluate new metadata schemes for their potential applicability (either directly or indirectly) in a library setting, we must

- examine our own functional requirements and have a clear sense of their relative priority (and expendability), and

- determine which of these requirements metadata constructed according to the new scheme will allow us to meet and which it will not.

Doing these things will enable us to make trade-offs in an informed way. What we might be trading away becomes clear when we look at another of Cilla's key points:

"Metadata schemes without content rules are not very useable"

Library cataloging according to AACR2 and MARC is the foundation of an incredibly complex and robust flow of data that libraries rely on not only for public access, but also for acquisitions, copy cataloging, resource sharing, cooperative collection development, and cooperative preservation. An amazing quantity of library metadata flows daily among countless computer systems in countless organizations. Our systems ingest this metadata, index it in sophisticated ways, sort it, identify duplicate records, correct errors in it, and update names and subject terminology within it. We can automate these functions precisely because the form and nature of the content is regulated and well-understood. Libraries have achieved an impressive degree of interoperability (that infamous word!) because we have created, maintain, and apply metadata standards, including cataloging rules. The initiatives Cilla describes have not yet resulted in anywhere near the volume of interoperable descriptive metadata that libraries have, but as they begin to scale up, they find themselves limited by the lack of consistency in their metadata.

Most of our catalogs, fortunately, still reflect Cutter's "objects", and we achieve these through the application of content rules and authority control. With consistently formulated metadata, we can provide fairly consistent and comprehensive retrieval of items by a given author, with a given title, and on a given topic. We can present searchers with well-ordered (and evidently ordered) lists of search results. In the absence of consistently formulated metadata we cannot do these things. To the extent that we choose to abandon or downplay content rules, we choose to limit what we can do, what functions we can support, with our metadata.

Martin Dillon calls for creating DC records in MARC.[1] Sarah Thomas calls for using Dublin Core in order to reduce the time spent cataloging books.[2] However, the decision to "use DC", whether in MARC or some other form, is not enough. It does restrict the universe of metadata elements, but was choosing which fields to fill in ever the tricky and time-consuming part of cataloging? It fails to address the question of how the content of the data elements should be selected, how it should be formulated, and whether any elements should be required. It is silent on the functional relationship of this metadata to other library metadata, particularly that in our OPACs. If we choose to create DC records without content rules or authority control, we must ask what the library can do with the resulting metadata. Can we continue to fulfill our objectives? Do we still have objectives, and can we articulate them at the level of detail that Cutter did?

A decision by libraries to use DC or any other metadata scheme as a native form of metadata should be accompanied or, ideally, preceded by far more important decisions about what functions this metadata must support and the rules that will be necessary to enable those functions. (And no, "catalog more stuff" does not constitute functional analysis.) By developing a library-community application profile for the Dublin Core Metadata Element Set, complete with refinements, extensions, and content rules, two things would become possible. DC would become a responsible option for native library metadata, and librarians could more fairly assess the costs and benefits of using DC as opposed to our current minimal or core-level cataloging records.

However, I am not advocating that the Dublin Core initiative as a whole create or adopt some cataloging code. Quite the contrary. I would modify Cilla's maxim slightly:

"Metadata schemes without content rules are not very useful as native forms of metadata."

The absence of content rules is critically important for two things the Dublin Core does extremely well: 1) serve as a kernel around which specific communities can develop their own richer metadata schemes, and 2) mediate between community-specific metadata schemes.

The Dublin Core was not intended to be sufficient to meet the internal metadata requirements of any single community. In fact, as a "native" scheme, it is almost always used with refinements and extensions. It is true that for each community, the process of developing content standards is difficult and time-consuming. However, given the variety of materials and perspectives represented in the DC effort today, the difficulty of creating or imposing a single content standard is immense. Nor I am convinced that these communities would continue to see Dublin Core as a viable option were they forced to adopt a single content standard.

Cilla has identified three ways the Dublin Core can be used to achieve the second goal, that is, to mediate between community-specific metadata schemes:

- as a minimal set of commonly understood access points for cross-domain searching,
- as a common extract format for creating union catalogs,
- as a searchable entry point to local files of more complex metadata

Note that each of these implies the existence of a richer description, presumably operating within an environment that takes advantage of that richness. The Dublin Core elements are, in these scenarios, either extracted from or mapped to elements in that richer description. If the Dublin Core Metadata Element Set itself were to be tied to a particular set of content rules, it would be difficult if not impossible to use it in these ways, since such Dublin Core-specific rules would inevitably conflict with community-specific rules. Which brings us to Cilla's next point:

"Simply mapping from a semantic or syntactical element in one scheme to a comparable element in another does not guarantee the usability of the converted metadata."

We have a tendency in this community to wave our hands and say "Oh, we have crosswalks -- it'll be fine." (And for use of Dublin Core as a so-called "switching language" among richer metadata sets, it generally is, because all we are trying to accomplish is fairly coarse, high-level discovery.) But the fact is that mapping between metadata schemes always results in loss: loss of data, loss of meaning, loss of specificity, loss of accuracy. As Caroline Arms notes in her paper, it is relatively easy to map from a richer scheme into a simpler one, accepting such loss, but mapping between rich metadata schemes is difficult, costly, and, I would add, rarely very effective.[3] What you get is often the proverbial dancing bear: it's not that he does it well-- the wonder is that he can do it at all. Or as Greg Colati of Tufts recently noted, mapping between metadata schemas enables us to communicate in grunts.

Semantic and syntactical mapping are themselves extremely imperfect. Differences in concept and in specificity inevitably result in metadata which reflects the lowest common denominator. In addition, as Cilla points out, metadata must be created according to content rules in order to be reliably useful. Therefore, any mapping which does not also transform the element content where applicable will result in metadata that is not very useful. The application of content rules usually requires human judgment in conjunction with an examination of the resource itself. In contrast, mapping metadata from one scheme to another is done algorithmically, takes place in the absence of the original resource, and is generally performed without human intervention. Transformation of the content of the elements (its selection or its form) is rarely attempted during such a process, and for good reason.

Specifically, mapping will not necessarily allow metadata that was designed to operate in one context -- in support of a given set of functionality -- to operate in another context, in support of a different set of functionality. Converted metadata will certainly not operate as well as metadata created expressly for the context would. This view contrasts with that expressed by Carl Lagoze in his paper. Carl advocates that libraries promote the catalog as a "mapping mechanism", and envisions an environment based on a "model that recognizes distinct entities that are common across virtually all descriptive schemas - people, places, creations, dates, and the like - and that includes events as first-class objects."[4] Carl is certainly correct that libraries have a mediating role, and high-level discovery across domains is a service many libraries are actively developing. Hopefully, these services will be in addition to full-featured catalogs tailored to particular communities and particular kinds of research, not in place of them. But the kind of cross-domain interoperation that Carl envisions supposes a degree of coordination and like-mindedness among all information producers that is hard to imagine, and a level of operational complexity beyond anything we know today. Even further, it underestimates the fundamental differences in worldview among various information-describing communities.

How well mapping will serve you depends on how far apart the schemes are in structure, semantics, and content rules, and on how much functionality from either the source or the target environment you need to retain. Mapping has its uses, but we need to recognize its limitations up front and not oversell its capabilities.

Cilla also exhorts the librarians to

"...begin thinking about basic bibliographic metadata as a commodity, produced and exchanged by a number of communities in order to serve a number of purposes."

Metadata created in the publishing arena is a natural fit with libraries. Clearly this is an area where active coordination and necessary compromise could yield real benefits. However, the differences in approach are not trivial. As Cilla points out, publishers are even more passionate than libraries are about controlling certain information such as the identification of rights holders. Unfortunately, given their interest in current materials, the publishing community's set of rights holders intersects but is not coextensive with the set of personal and corporate names so important to libraries, which cover many centuries of authorship.

Conceivably, commodity metadata could extend into other arenas as well. Perhaps there is a role for commodity metadata in any cases where the thing described is mass-produced, mass-accessible, or mass-referenced. This takes us beyond bibliographic metadata for published materials and into metadata for commonly referenced but singularly occurring works such as the Mona Lisa or the Vietnam Veterans' Memorial, into metadata which describes agents such as authors, performers, etc., and into gazetteer-type metadata for geographic locations.

The examples Cilla gives of how we can fruitfully approach the meaningful sharing of metadata between the publishing and library communities:

- common authority lists
- shared content rules
- elimination of content rules that do not make enough of a functional difference to insist upon
- adoption of additional metadata elements to enhance the functionality we provide

are useful areas to examine whenever libraries want to interact with another metadata-producing community. Making an informed choice about which metadata schemes to adopt or to adapt requires analysis and decision-making at this level of detail.

Finally, Cilla makes a critical observation that is often lost in the frenzy to "catalog the web":

"The key question ...is not bibliographic control of Web resources, but rather bibliographic control of both digital and non-digital resources in the Web environment."

The theme of "selection" came up time after time in this conference. Libraries have never cataloged every take-out menu and place mat, which is the level of much of the content of the "free web" today. As more of the substantive content that libraries have always chosen to provide and preserve moves to restricted, fee-based web delivery, the more important our formal relationships with the publishing community will become, and with them, the ability to repurpose the "commodity metadata" of which Cilla has spoken.

It is no accident that most of the metadata schemes Cilla enumerated apply to both digital and non-digital media. Enormous quantities of valuable physical resources exist and will continue to exist, and any model of bibliographic control for the new millennium must take these into account. The web permits vast amounts of non-digital information to be exposed for discovery through descriptive databases. Such non-digital material exists not only in libraries but also in such diverse organizations as museums, natural history collections, and visual resource collections. Much of this information will never be made digital due to the impossibility of capturing its artifactual value in digital form, to the economics of converting and maintaining information in digital form, or to other constraints. The problem facing libraries derives only in part from the proliferation and complexity of web resources. It also lies in the challenge and opportunity to help users make sense of the flood of resources, both digital and non-digital, that the web reveals.

---

1. Dillon, Martin. "Metadata for Web Resources: How Metadata Works on the Web", Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000
2. Thomas, Sarah. "The Catalog as Portal to the Internet", Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000
3. Arms, Caroline. "Some Observations on Metadata and Digital Libraries", Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000
4. Lagoze, Carl. "Business Unusual: How 'Event-Awareness' May Breathe Life Into the Catalog", Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000

Library of Congress
December 21, 2000
Comments: lcweb@loc.gov

Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books?

A Reference Librarian's Thoughts on the Future of Bibliographic Control

Thomas Mann
Reference Librarian
Library of Congress
tman@loc.gov

Final version

Summary

Precoordination of LCSH subject headings, both (partially) in the LCSH thesaurus and (more extensively) in OPAC browse displays, continues to be necessary for several reasons:

- The meaning of thousands of LCSH headings depends on their word order in ways that cannot be captured by postcoordinate Boolean combinations or by word proximity searches that drop relational prepositions as stop words.
- A vast network of linkages between LCSH headings and the LCC classification scheme depends on precoordination i.e., changes in the word order of the subject strings also changes the classification areas to which the terms point.
- Displays of precoordinated strings enable researchers to simply recognize whole arrays of relevant research options that they could never specify in advance in postcoordinate combinations. The larger the file, the more such recognition capabilities are necessary.
- The precoordination of terms is inseparably linked to a vast network of cross- references that

would vanish without it.

Books are not vanishing or generally evolving into digital forms; they continue to be published in huge numbers every year, and they provide formats that are more readable for lengthy texts.

In the future, LCSH must serve in both the environments of online library catalogs and the Web not the latter in place of the former.

An Online CIP (OCIP) program would enable our profession to maintain the necessary precoordination of LCSH headings in OPACs and also to insert librarian-created LCSH elements into the Web headers of participating online publishers. This would enable us to exploit the existing precoordination and postcoordination capacities of OPACs, and also to exploit LCSH more extensively in the exclusively postcoordinate search environment of the Web.

LCSH headings in copy cataloging cannot be simply accepted "with little or no modification."

Full text of this paper is available in PDF format.

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books?
A Reference Librarian's Thoughts on the Future of Bibliographic Control

Thomas Mann

Reference Librarian
Library of Congress
tman@loc.gov

Summary

Precoordination of LCSH subject headings, both (partially) in the LCSH thesaurus and (more extensively) in OPAC browse displays, continues to be necessary for several reasons:

- The meaning of thousands of LCSH headings depends on their word order in ways that cannot be captured by postcoordinate Boolean combinations *or* by word proximity searches that drop relational prepositions as stop words.
- A vast network of linkages between LCSH headings and the LCC classification scheme depends on precoordination–i.e., changes in the word order of the subject strings also changes the classification areas to which the terms point.
- Displays of precoordinated strings enable researchers to simply *recognize* whole arrays of relevant research options that they could never specify in advance in postcoordinate combinations. The larger the file, the more such recognition capabilities are necessary.
- The precoordination of terms is inseparably linked to a vast network of cross-references that would vanish without it.

Books are not vanishing or generally evolving into digital forms; they continue to be published in huge numbers every year, and they provide formats that are more readable for lengthy texts.

In the future, LCSH must serve in both the environments of online library catalogs and the Web–not the latter in place of the former.

An Online CIP (OCIP) program would enable our profession to maintain the necessary precoordination of LCSH headings in OPACs *and also* to insert librarian-created LCSH elements into the Web headers of participating online publishers. This would enable us to exploit the existing precoordination and postcoordination capacities of OPACs, and also to exploit LCSH more extensively in the exclusively postcoordinate search environment of the Web.

LCSH headings in copy cataloging cannot be simply accepted "with little or no modification."

Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books?

A Reference Librarian's Thoughts on the Future of Bibliographic Control

Thomas Mann

Aristotle wrote that "The least initial deviation from the truth is multiplied later a thousandfold"; Mortimer Adler similarly paraphrases Thomas Aquinas in saying "little errors in the beginning lead to serious consequences in the end."[4] The point here is that participants in this Conference need to pay particular attention to initial, unargued assumptions about the very purposes of cataloging and metadata if we wish to ward off some very large unintended, but nonetheless very undesirable, consequences if those purposes are inadequately assessed right at the beginning.

My major concern is this: Some of the papers before this Conference suggest that the Library of Congress Subject Headings system (LCSH) can be tailored to the task of Web cataloging by eliminating–or at least substantially reducing–its precoordinate displays of subject strings, both within the basic list itself and within browse displays in online catalogs. There is even a suggestion that such browse displays of strings of terms are entirely unnecessary, given the computer's ability to do postcoordinate Boolean combinations. I will demonstrate in some detail that this belief–often apparently more assumed than forthrightly stated–is extraordinarily naive. If, as a result of this Conference, the researchers of this country lose precoordinated displays of terms in LCSH–which serve several definite functions that are apparently being overlooked–then future scholars will have much less efficient subject access to large book collections. The gains–if they come about–achieved in better access to Web sites will be more than vitiated if they are accomplished at the expense of losing access to large (and still growing) book collections by undercutting the many functions of LCSH that require precoordination.

One immediate recommendation

Before examining what I think are bad ideas, let me jump ahead to one recommendation that I hope this conference will consider. As a reference librarian I'd very much like to see browse displays like this in catalogs of the future, integrating references to both books and Web sites:

Women–Services for
Women–Services for–Bolivia–Directories
Women–Services for–Caribbean area–Case studies
Women–Service for–Ethiopia–Congresses
Women–Services for–Germany–History

Women–Services for–Michigan–Evaluation
Women–Services for–New Zealand–Bibliography
Women–Services for–North Carolina–Finance
Women–Services for–Study and teaching–United States
Women–Services for–Study and teaching–United States–Web sites (.edu)
Women–Services for–United States–Directories
Women–Services for–United States–Web sites (.com)
Women–Services for–United States–Web sites (.edu)
Women–Services for–United States–Web sites (.edu)–Data archives
[This "Data archives" subdivision may not be appropriate for this particular subject; I offer it here just as a pattern example.]
Women–Services for–United States–Web sites (.edu)–Discussion lists
Women–Services for–United States–Web sites (.edu)–Portals
[I'm using "–Portals" here; "–Site directories" might be an alternative, in which case a cross-reference is needed: Site directories USE Portals]
Women–Services for–United States–Web sites (.gov)
Women–Services for–United States–Web sites (.org)
Women–Services for–Wisconsin–Periodicals
Women–Services for–Zambia–Directories

Such a display would enable researchers to *recognize* selected, high quality Web sites *in relationship to* the substantive knowledge records in the library's book collections–which are not, and for the most part never will be, digitized. (Of course there should be live links from the catalog records to the Web sites insofar as licensing agreements allow.)

In contrast, reliance on exclusively postcoordinate combinations such as Women AND Services AND "Web sites" would conceal the *relationship* of the Web resources to the relevant books.

Both precoordination and postcoordination necessary

The presence of a such a precoordinated browse display, of course, does not preclude postcoordinate Boolean search capabilities. Neither I nor anyone else is arguing for precoordination *rather than* postcoordination. We need *both* browse displays of precoordinated strings *and* the possibility of postcoordinate combinations of individual elements.

Browse displays, above all, enable us to recognize search options that we could never specify in advance, in Boolean combinations, by showing them to us in relation to options that we can think of. The larger the file, the more researchers (and reference librarians) need this recognition capability. What I am afraid of is the dismissal, on inadequate grounds, of the continuing importance of browse displays of ordered subject strings.

The loss of precoordination in LCSH in the Web/networked environment would cause very serious retrieval problems if the same loss were extended to LCSH in the OPAC environment. Since there's no point in maintaining two different LCSH systems, these very real problems in the OPAC environment have to serve as a brake on the otherwise free-floating speculations, untied to real library collections, that inform many of the projections of LCSH's future when considered exclusively in the Web environment.

When speaking of precoordination in LCSH, we must distinguish two different locales in which subject phrases must be displayed, although to varying degrees: first, within the LCSH list itself; and second, within online catalog browse displays, which show the linkage of free-floating subdivisions to headings, not displayed in the list itself.

Meanings of LCSH terms and links to LCC dependent on precoordinated word order

The first reason that precoordination must continue to be shown in the LCSH list itself lies in the need to capture intellectual meanings dependent on word order or prepositional relationships that are not captured by postcoordinate Boolean combination, or by simple word-proximity searching.

Moreover, such ordered combinations often entail specific links to the classification scheme. The order of the words in the headings changes the Library of Congress Classification (LCC) areas to which the headings are linked.

For example, the string **Philosophy–History** is spelled out precoordinately in the LCSH list even though "History" is elsewhere a free-floating subdivision. Why does the relationship of these terms need to be spelled out like this? and why does it then need to be precoordinated *in the LCSH list* rather than simply *in the catalog's browse display* of other subdivisions under "Philosophy"?

The phrase needs to be precoordinated to begin with because the order of the terms changes the meaning of the phrase: **Philosophy–History** is not the same thing as **History–Philosophy**. The phrases need to be combined *in the list* because additional information about the subjects must also be conveyed to both catalogers and catalog users: that *a change in the order of the terms also signifies a change in the classification areas appropriate to the different phrases*:

- **Philosophy-History** is explicitly linked to a major clustering of books on this subject in the B69-B4695 areas of the classified bookstacks.

- **History–Philosophy**, on the other hand, is explicitly linked to the D16.7-D16.9 areas of the stacks.

This explicit linkage of LCSH to the Library of Congress Classification scheme (LCC) permeates the length and breadth of the subject heading list. (This important fact is simply overlooked in some of the papers before this Conference. It is perhaps noteworthy that the *Sears List of Subject Heading* is linked to the DDC system in just the same way.)

A postcoordinate combination of History AND Philosophy (in Voyager, entered as +history +philosophy in the keyword search mode) will, first, exceed the system's display limit of 10,000 records in my library's catalog. Second, the display of the 10,000 records that *are* retrieved will show in its first fifteen items–the ones that are most highly "relevance ranked"–classes numbers scattered among B, BD, DA, DT, GV, HC, HG, JN, ML, PA, QA, and Z. Not only does the meaning of the words change when their precoordinate ordering is lost; the specific areas of the bookstacks most closely associated with those different meanings are also concealed from a researcher's view.

If the two terms, **Philosophy** and **History**, are searched not as keywords but as subject terms confined to the controlled 6XX subject fields, their postcoordinate combination will *still* produce (in my library's catalog) a retrieval in excess of the 10,000 records that can be displayed; and the first twenty that do show up and have class numbers are scattered among AS, B, BH, BQ, CB, GV, H, HV, LA, PQ, and Q areas. The reader will be overwhelmed with "relevance ranked" junk, and will also be prevented from knowing which stack areas would be best to browse for full-text information.[2]

Even faceted elements must sometimes be displayed in precoordinated strings

Even if there is, quite properly and usefully, much faceting in LCSH so that the same subdivision can be applied to many headings, the *display* of *some* heading-subdivision combinations must still be shown in precoordinated manner in the basic list. This is because the order of the words is often tied to particular classification "cluster" areas. Another example is the heading **Women–Services for**, which in our catalog (including all further subdivisions) turns up 176 records, with noticeable clustering of the referenced books in three class areas, HV1442-1448, HQ1236.5-1240, and the HQ1740s.

A relevance-ranked keyword search of Women AND Services (in Voyager, +women +services), however, turns up and overwhelming 1,797 records (of which 1600 are books). Of the "most relevant" fifteen displayed first, only two records show up in any of these three clusters, and in two separate ones at that (i.e., one gets a sense only of individual items, not of important clusters). In other words, the "relevance" ranking completely erases from a searcher's perception the existence of such aggregates in the bookstacks–groups of related books, shelved together, that *are* brought to his attention via the precoordinated subject strings.

Additional linkages between LCSH strings and LCC show up in the catalog, not in the thesaurus

In this case, it is noteworthy that the **Women–Services for** heading is, within the LCSH list itself, explicitly linked only to HV1442-HV1448--but in the library's actual catalog, a search under this string will bring up records that show definite clustering in the two additional areas just mentioned. In other words, the linkage of LCSH to the classification scheme is by no means simply a "one to one" connection. Its full complexity is discovered only by actually searching the precoordinated headings *in the actual catalog*, at which point the retrieval of records under the various subject terms may indicate yet other important clusterings associated with a particular string–which clustering areas are not formally indicated by LCSH-LCC links within the thesaurus itself.

This may sound sloppy to theorists who don't use actual catalogs and bookstacks very often; but my own experience is that the many linkages just *work*. The relationship of LCSH and LCC is partly specifiable in the LCSH list; but, in large part, the full extent of the interconnectedness of LCSH and LCC is discoverable only in the library catalog itself. This network of interconnections probably defies fully coherent *a priori* specification; but it nonetheless *functions* in the real world to direct readers from headings in the catalog to particular areas in the stacks. I sometimes think of New York City's underground as an analogy–the intertwinings of water lines, sewer tunnels, heating ducts, and electrical and optical conduits probably cannot be full determined on an *a priori* basis simply by looking at a blueprint or schematic (analogous to the LCSH list); one has to actually go down into a manhole to grasp fully what's wrapped around what (analogous to the full catalog). The larger point, however, is that we naively tamper with such myriad interconnections at our peril–and we certainly shouldn't embark on such a course by naively *overlooking the very existence of these linkages in the first place.*

Another analogy would be that of language: language does not fully reduce itself to neat rules that can be specified *a priori*. It develops on its own, in ways that defy logic. Just so is the relationship of all of the LCSH-ed records in a library catalog to all of the LCC-classified books in the stacks: the former definitely point to the latter, but logical rules spelled out beforehand are not always the best guide to the connection. Over the course of a century, the connections "just growed." To pretend that they are not there, however, and to simply ignore the continuing need for the catalog's precoordinated headings to point to particular "clustering areas" in the classified areas of the bookstacks, would be to do enormous harm to our nation's research libraries.

Additional examples of term meanings and links to LCC dependent on precoordination

Postcoordination of the terms, then–if relied on as the sole means of subject searching–utterly destroys not only the meanings of different subjects that contain the same words, but also the indexing of the class scheme that takes place when the subject terms are displayed in meaningful precoordinate relationship-strings. *A change in the order of the words also entails a change in the classification areas.* Other examples:

Indian women is not the same as Indian AND Women
    Indian women–Mexico is linked to F1219.3.W6

Indian Women–North America is linked to E98.W8
Indian Women–South America is linked to F2230.1.W6[3]

Jewish women (linked to HQ1172) is not the same as Jewish AND Women

Women alcoholics (linked to HV5137) is not the same as Women AND Alcoholics

Women clergy (linked to BV676) is not the same as Women AND Clergy

For the sake of researchers who continue to use the bookstacks of major American libraries–and especially for the sake of the advanced academics in a wide variety of disciplines who are not represented at this Conference–we cannot naively overlook this extraordinary web of relationships linking these phrases (both in the LCSH list and in actual catalogs using LCSH) to the classification scheme.

A searcher who makes use of the precoordinated headings will thus be given important "focusing" information regarding which areas of the stacks to go to for the best groupings of knowledge records–*books*–for in-depth searching of *full-texts, back-of-the-book indexes, and prefaces* relevant to her topic–which knowledge elements are *not in the OPAC or on the Web*. A searcher who relies on postcoordination of separate elements will be overwhelmed with junk, and, further, will have no idea which stack areas would be best to examine first.[4]

Precoordination needed to capture prepositional relationships

Other terms need to be precoordinated in LCSH because *prepositional relationships* are crucial to the meaning of the terms–and prepositions vanish as stopwords in both postcoordinate Boolean combinations and word-proximity searches.

For example, searchers who browse **Women *on* television** will find 53 titles and be pointed, in LC's catalog, to particular *clusters* in PN1992.8.W65 and PN1995.9.W6. Searchers who browse "Women *in* television" will find the heading **Women in television broadcasting**, which will identify a third cluster of records at HD6073.T382 (Classes of labor. Women. Special industries or trades). Only one book–not a cluster–in this HD area shows up under "Women *on*" rather than "Women *in*" television in LC's catalog.

Researchers who simply use the keyword "relevance ranking" software will, in combining Women AND Television (in Voyager, +women +television) will be inundated with 804 records, only 345 of which are book records; and of the top twenty "relevance ranked" records (disregarding unavailable in-process or incomplete CIP records), *none* fall into any of these three most-relevant clusters in the bookstacks. The indexing function that the catalog serves in relation to the classification scheme is utterly lost without precoordination.

Once again, postcoordination of separate words effectively erases important information linked to the precoordinated term-order in the subject heading. From the existing browse displays of the ordered subject strings, however, researchers are effectively guided to go *here*, *here*, and *here* for the best groupings of in-depth (full text) knowledge records in the bookstacks. Without such direction to the stacks provided by precoordination in LCSH, researchers in this country will have a much more difficult time finding substantive knowledge records–books–in libraries.

Additional examples of prepositional relationships requiring precoordination

Other examples of prepositional relationships and indexing information that would be lost without precoordination:

Motion pictures for women (linked to PN1995.9.W6) is not the same as Motion pictures AND Women

Photography of women (linked to TR681.W6) is not the same as Photography AND Women

Sexual ethics for women is not the same as Sexual ethics AND Women

Social work with women is not the same as Social Work AND Women

Violence in women is not the same as Violence AND Women

Women, Black, in art is not the same as Women AND Black AND Art

Women in advertising is not the same as Women AND Advertising

Women in art (linked to N7629-N7639) is not the same as Women AND Art

Women in communication (linked to P96.W6) is not the same as either **Women–Communication** or Women AND Communication

Women in development (linked to HQ1240) is not the same as Women AND Development

Women in the Bible (linked to BS57.5) is not the same as Women AND Bible

Women in Church work is linked to BV4415

Church work with women is linked to BV 4445

Church work with women–Catholic Church is linked to BX2347.W6

If we do not maintain such precoordinated displays in LCSH and in catalog browse displays, this Conference will be seriously crippling the field of Women's studies–we will be making it much more difficult for scholars in this area not just to find, but to get a *structured overview* of books relevant to their topic within research libraries.

The Goal of Cataloging

Let's keep in mind that the goal of *cataloging* is not simply to give researchers "something." That goal can nowadays be accomplished by simple keyword searching without any intelligent human intervention in the forms of categorization, standardization of terminology, linkage of disparate concepts, and structured displays of search options. The goal of cataloging, in contrast, is to give researchers ***an overview of the extent of the relevant resources*** available for their topics (this is a year 2000 paraphrase of the intent of Cutter's "what the library has"). Overviews require connections, cross-references, and displays of options that cannot be specified in advance by researchers who literally don't know the fields they're getting into, and who often barely know how to phrase their initial questions. Overviews require displays of *relationships*, not just isolated data. These cannot be achieved without some measure of precoordination.

"Heavy lifting" capability required in research libraries

I realize that maintenance of precoordination makes LCSH more complex than it would be if it were simply an entirely faceted system of individual elements available for postcoordinate Boolean combinations or word-proximity searches. But complexity is sometimes simply necessary in order to get important jobs done. The control panel of a giant C5-A transport plane is necessarily much more complex than that of a Piper Cub twin-seater. If the Air Force were to reduce the former to the simplicity of the latter, they would soon find that their major "heavy lift" vehicle is capable of transporting materiel only by taxiing along the ground for short distances instead of flying with heavy loads over long distances. They would lose their ability to *lift* heavy loads into the air.

In a similar way, research libraries have to maintain their "heavy lifting" capacities with their unparalleled local resources, inside their walls. (It is especially the "heavy lifting" capacities that United States libraries have in providing *subject access* to their collections that make them the envy of other libraries–and scholars–throughout the rest of the world.) Granted, not every researcher needs the full capacities of the retrieval system for every inquiry. But the full capacities still have to be maintained for the frequent and unpredictable times when they are needed. To return to the plane analogy, our country doesn't need a C5-A every time a package needs to be delivered; but it does need the C5-A to be in readiness at a moment's notice.

Isn't the level of our intellectual research capacity–*which is our profession's* *responsibility*–just as important to this country as its military capacity? My experience as a reference librarian is that even questions that initially sound very "simple," from ordinary citizens rather than advanced scholars, often have a way of quickly escalating into inquiries that do indeed require the "heavy lifting" capacities of libraries. Whenever that happens we must be able to respond with more than just "something." We need to be able to map our way efficiently into the range of knowledge records available, not just respond with isolated information.

If we as professionals are not making knowledge available–in its largest possible frameworks of relationships, interconnections, and linkages–rather than just isolated bits of information, then we are nothing at all. If we see ourselves as providing access only to information rather than knowledge, or to information as a higher priority than knowledge, then we can indeed be replaced by machines.

Effects on Women's studies and Black studies

If we throw away precoordination in LCSH–which gives us so much of our "heavy lifting" capacity–we will be crippling not just the field of Women's studies, but that of Black studies: the arrays of precoordinated headings starting with the term **Afro-American(s)**–apparently soon to be changed to **African American(s)**–is fully as complex as the array of **Women** headings. I urge everyone participating in this Conference to take a look at the red books' thirty-five columns of precoordinated **Afro-** phrase headings arrayed on twelve pages–and this even without free-floating subdivisions being fully displayed.

Several times I have helped students who came in saying that they had to write a paper on "Black history." By alerting them to the amazing bounty of options they never knew they had, spelled out for their simple recognition just within the LCSH list (let alone within the catalog's browse display), such students are enabled to focus their topics in a wide variety of ways that would simply not otherwise occur to them. **Afro-American healers**, **Afro-American pacifists**, **Afro-American outlaws**, **Afro-American orchestral musicians**, and **Afro-American whalers** are all part of Black history; and these are only a very tiny sampling of the hundreds of options that would simply vanish from the radar screen if the searchers tried only **Afro-Americans** AND **History**.

Giving researchers *overviews of what is available*–opening up their eyes to unsuspected possibilities, positioning them on conceptual maps of options, and anchoring them within relevant intellectual frameworks–*this* is what public service is about; it is not a matter of giving them simply "something."

The **Afro-** headings, too–just as with the **Women** headings–tie particular aspects of Black studies embodied in precoordinated phrases to widely different areas of the classification scheme. For a five-page example of this point–which I mercifully will not reproduce here–see my *Library Research Models* book (Oxford U. Press, 1993), pp. 33-37.[5] If we unwittingly destroy the precoordinated display of the **Afro-** headings we will simply decimate the research potential of Black studies scholars in American libraries.

How would such a development be reported in *The Chronicle of Higher Education* or *Lingua Franca*? (What would Nicholson Baker have to say about it in *The New Yorker*?!) Would it reflect credit on us? Or would it show that, in order to remedy our whole profession's traditional inferiority complex, we sold subject access to book collections down the river in order to appear more "with it" in Web searching?–and did so with the full knowledge that, while librarians can reasonably structure access to book collections in research libraries, we will never be able to intelligently apply LCSH to more than a microscopic sampling of the *billions* of Internet sites available. Will it be reported that we gutted precisely the elements of LCSH that make it so useful *in* structuring access to book collections, in order to facilitate unstructured applications of individual terms (stripped of both their contextual strings and links to LCC) to Web site records? Why do we assume, in the first place, that anyone will turn to *library catalogs* for primary access to the Web when field is already taken by Google, AltaVista, NorthernLight, Hotbot and a dozen other more comprehensive search tools?

The Virtue of OPAC Coverage of Web Sites

If library catalogs are to cover Web sites–and indeed they should, selectively–then their virtue will be precisely in bringing Web sites *into relationships* with the *substantive knowledge records* that *books* are–especially since book collections, for copyright and preservation reasons alone, will always reside primarily off the Web, within library walls. We need to tie the two sources together, not sacrifice one to the other. And one part of the linkage of the two environments–another will be discussed below–will be brought about most effectively by extending rather than eliminating the range of our precoordinated browse displays in our catalogs, as in the **Women–Services for** example above.

Precoordinated Word Order Also Affects Cross-Reference Structure

There is yet another reason not to destroy the display of precoordinated strings in LCSH: not only does the meaning of subject terms change depending on the order of their words; not only does the huge web of linkages between LCSH and LCC depend on the word-order of the terms; not only do the meanings of proximate nouns in the same order need to be distinguished by different prepositional relationships–not only for all of these reasons does precoordination need to be maintained in the OPAC environment, but for another reason, too: the order of terms also *critically affects the cross-reference structure* between and among related terms. (Of course cross-references don't show up in Web-type searches–the software can't handle them. Does that mean that they're now *also* dispensable in the OPAC environment?) Let me give just two examples from the hundreds of thousands available:

The precoordinated phrase **Women-Psychology** (which is explicitly tied to HQ1206-HQ1216 in LCC) is linked by cross-references to:

RT Women–Mental health
NT Achievement motivation in women
  Animus (Psychology)
  Anxiety in women
  Assertiveness in women
  Body image in women
  Cooperativeness in women
  Helplessness (Psychology) in women
  Leadership in women
  Self-esteem in women
  Self-perception in women[6]

This entire network of relationships would be lost if users could search only Women AND Psychology. Researchers could find only isolated information, not a web of knowledge relationships.

The precoordinated phrase **Afro-Americans–Education** (which is explicitly tied to LC2701-LC2853 in LCC) is linked by cross-references to:

BT Education–United States
RT School integration–United States
NT Afro-American students
  Afro-American women–Education
  Afro-Americans–Professional education
  Afro-Americans–Scholarships, fellowships, etc.
  Afro-Americans–Vocational education
  English language–Study and teaching–Afro-American students
  Segregation in education–United States
  Segregation in higher education–United States

Once again, all of these displayed linkages that bring to researchers attention options they would not otherwise perceive–all would be lost if, in order to make LCSH more "flexible" for a Web environment, we throw away precoordination in the OPAC environment. (Do we really want to do this? As the kids these days say, Isn't this a "no brainer"?)

Key Functions of LCSH Being Overlooked

Unfortunately, none of these problems entailed by eliminating precoordination are even mentioned by key papers before this Conference. (Even beyond this meeting, there are many cataloging theorists out there who seem to think that the *only* function of precoordination is "to break up large files." Where do they acquire such blinders? Is this what is being taught in schools of library and information science? Perhaps less time in the academic ivory tower and more time working behind public service desks in real libraries is indicated.)

Let me turn to several other assumptions that show up in some of the papers–all of which affect the precoordination/postcoordination issue–and that I think are "not ready for prime time."

Information and Knowledge Are Not the Same

The first of these is something to which I've already alluded. It is the assumption that *information* and *knowledge* are the same thing, and can be formally handled by retrieval systems in just the same way. I beg to differ.

First, there is a real hierarchy in the realm of human awareness. The lowest level is formed by *data*, the unorganized, unfiltered, and unevaluated raw material of thought, comparable to sense experience (although, I think, not reducible to it–but that's another paper). *Information* is at a higher level, reflecting an organization of data to the point that statements can be made about it, either true or false, and coherent or incoherent with other information. *Knowledge* reflects a still higher level of organization to the point that truth or falsity can be reasonably assured by tests of correspondence to, and coherence with, the world of experience and of other ideas; it requires that information be put into much larger frameworks of relation to the worlds of matter and ideas. This level includes discernment of patterns and *interconnectivities* within information, and the making of generalizations that are accessible to, and acceptable by, other people. (I won't belabor here the further levels of understanding and wisdom.)

Information simply does not have the degree of "truth-claim" upon us that knowledge has, because it does not have the *connectedness* and *relatedness* of knowledge; and, further, it also depends on all of the larger frameworks of knowledge, understanding, and wisdom for an assessment of its *worth*.

These are not merely academic distinctions; they have a material bearing on the very purposes, methods, and materials of *cataloging* and bibliographic control.

Conveying Knowledge Requires Larger Cataloging Structures and Linkages

Briefly: We ought not to be dismantling the larger structures and webs of *knowledge* that cataloging has created in order simply to achieve less costly access to *unintegrated information*. Access to *information* is much more amenable to automatic machine methods of indexing, without human structuring, than is access to *knowledge*; but automatic methods of gaining access to information are not sufficient to show researchers the *knowledge relationships embedded within LCSH subject-strings themselves, within their cross-references, and within their integral connections to the Library of Congress Classification (LCC) scheme.*

Screen Displays and Book Displays Change Readability

The next assumption that we need to examine is the assertion that knowledge is equally well conveyed by screen displays as by book formats. I doubt this very much. How many of us are now reading book length narrative or expository works–say, the equivalent of a 200-page book–on screen displays? I'm not talking about long lists of hits on Google or Yahoo, or long lists of directory information, or bibliographical listings, or long rosters in Ebay; I'm talking about long, coherent narrative or expository texts. Some are reading such things on screens, I'm sure; but I'll just remind everyone to examine his/her own reading habits before imposing theoretical projections upon everyone else. If we don't read long connected texts on screen displays ourselves, let's not force others to be shunted by our catalogs exclusively or even primarily to Web sites rather than printed books.

Knowledge–requiring longer attention spans to establish its connectedness–is much more readily conveyed by *book formats* than by screen displays of textual material, which most people recognize as being "slanted" to shorter attention spans.[7] If this is true–and I think it is–then this Conference should not cavalierly assume that future catalogs ought to be more concerned with Web sites than with books. Catalogs need to cover both–but not the former in preference to the latter. Let's not forget, right at the outset, that book formats are a proven medium for conveying knowledge, while the verdict on Web sites is truly not yet in–and may not be as rosy as some are assuming. (The additional problem of changing the focus of library catalogs from books to Web sites is that of *preservation*–it is neither inevitable nor even likely that electronic resources can be preserved at nearly the cost-efficiency of preserving books.)

I strongly agree with Walt Crawford and Michael Gorman's initial position in their book *Future Libraries: Dreams, Madness, and Reality*: "Let us state, as strongly as we can, that libraries **are not wholly or even primarily about information**. They are about the preservation, dissemination, and use of recorded knowledge in whatever form it may come . . . so that humankind may become more knowledgeable; through knowledge reach understanding; and, as an ultimate goal, achieve wisdom."[8]

The *book format* is by far the best vehicle that humanity has devised for conveying to itself the higher levels of knowledge and understanding, and the *research library* is the best vehicle that has ever been devised for making large collections of substantive knowledge records *freely* available, without prohibitive individual subscription costs or point-of-use charges, or on-the-spot printing charges. Most of the billion+ Web sites, of course, are not substantive; and a high percentage of those that are most desirable are generally confined by license agreements to particular terminals within walls, or to tightly-defined user groups–i.e., such sites *cannot* be tapped into freely by anyone, from anywhere, at any time. In that sense they are much like books: *freely* available only *within library walls.*

Library Catalogs Provide *Alternatives* to the Web

Library catalogs, if they are to have an important function in the age of Google, Altavista, and NorthernLight, would serve users best by directing them to selected, substantive sources of knowledge–especially to the abundance of sources that are not, and never will be, freely available to anyone, from anywhere on the Web. This means that catalogs will function best by presenting researchers not just with different ways to search the Web, but with substantive *alternatives to* the Web, especially copyrighted or licensed resources that cannot be found within the vast ranges of free Web sites. (Most users think of "the Web" as the free portions of it; I find this repeatedly when I show researchers our licensed databases–their question is always phrased as "Can I get this on the Web?," but their meaning is "Can I tap into this for free outside the library walls?")

Other Questionable Assumptions

Beyond the misleading assumption that information and knowledge are the same, there are other questionable assumptions that we need to be on our guard to spot, all of which may be found in current literature, and some of which show up in some of the papers before this Conference:

- that "knowledge" records, in general, are now making a "transition" to digital forms;
- that the *only* context in which we must regard the future of bibliographic control is one of shared Web access–i.e., that the context of continuously expanding and *localized book collections* need no longer concern us as a higher priority;
- that the functions of cataloging in the persisting *book collections context* can now be dispensed with–without even examining what those function are–insofar as they are not readily adaptable to the context of accessing Web sites;
- that, specifically, precoordination in displays of LCSH subject heading strings is no longer necessary either as (partially) enumerated in the LCSH list itself; or as (fully) displayed in "browse"screens in online catalogs, because postcoordination of individual elements renders such string-displays intellectually "unnecessary" or, worse, socially stigmatizes them as "old fashioned" (thereby precluding any objective assessment of their continuing functions)
- that researchers of the new millennium will choose *library catalogs*, to begin with, as their *primary* avenues of access to the Internet;
- that library catalogs, preeminently, must dominate the information landscape of the future by "seamlessly" leading researchers to *all* of the information they may need (rather than serving more modestly as one channel of access to *some*

important knowledge and information records).

- that catalogs will, ought to be, and *can* be used successfully–i.e., to give inquirers an overview of their research options and to lead them to the best information/knowledge on their subjects–by *untrained* researchers *in isolation*, that is, in the absence of any intervention by reference (or other) librarians, either beforehand in bibliographic instruction classes, or immediately at the point of use. (This would be analogous to Piper Cub pilots trying to fly C5-A transports, with their much more complex control panels, without any help.)

- that, rather than using catalogs to integrate the two contexts of knowledge records contained in books and substantive Web sites, catalogers of the future should markedly diminish their concern for books and concentrate on Web sites instead.

- that any concern for maintaining precoordination in LCSH should be dismissed *a priori* on the grounds that, because it first developed within manual catalogs, precoordination is a mark of outdated, "pre-high-tech" thinking. (This is nonsense. Precoordination makes *online catalogs* function much more efficiently.)

Are Books Evolving into Digital Forms?

Martin Dillon, in a (thankfully) "blunt statement," works from one initial assumption very different from my own:

> After a long and various evolution, knowledge representation settled into paper products for most of its output. Now we are shifting to digital forms for representing knowledge and to the Web as the primary distribution channel. This change will have profound consequences. There is little question, for example, that paper products will gradually be replaced by Web-accessible digital products.[9]

I respectfully beg to differ. Even F. W. Lancaster now has "Second Thoughts on the Paperless Society."[10] Walt Crawford, in his article "Paper Persists: Why Physical Library Collections Still Matter,"[11] makes a number of relevant points:

> What happens if the premises arguing for library conversion to digital fail? Logically, if the premises are invalid, then the conclusion is false or at least unsupported.
>
> * * *
>
> Reading from digital devices, whether portable or desktop, suffers in several areas–among

them light, resolution, speed, and impact on the reader–and there has been essentially no improvement in any of these areas in the last five years.

Many futurists have conceded this point. They now admit that people will print out anything longer than 500 words or so. It's just too hard to read from a computer, and it doesn't seem likely to get a lot easier. If every long text is printed out each time it is used, there are enormous economic and ecological disadvantages to the all-digital library: briefly, a typical public library would spend much more on printing and licenses than its current total budget and would use at least 50 times as much paper as at present.

Continuing Production of Book Formats in Huge Numbers

It is also worth noting that the new *Bowker Annual* (2000) has, just this year, radically revised upward its statistics on the number of books produced in this country in recent years; last year it recorded 1997 book title production as 65,769 titles; now it records 1997 production as 119,262 titles. Similarly the revision of the 1998 figure is from 56,129 to 120,244 titles. It seems more than questionable to assume that *books* are making "the transition" that is so cavalierly assumed in so much information science literature these days. Research libraries are still heavily anchored in print collections *as well as* in digital resources; and the latter simply are not the only context in which LCSH must function.

Significant Differences Between OPAC Cataloging and Web Metadata: Displays of Relationships

Mr. Dillon makes a further point, with which I do not disagree, in quoting a description of metadata:

> *Meta-information has two main functions:*
> - *to provide a means to discover that the data set exists and how it might be obtained or accessed; and*
> - *to document the content, quality, and features of a data set, indicating its fitness for use.*[12] [italics in original]

This is fine–as far as it goes. But *cataloging*, unlike metadata, has additional functions beyond these two, especially in the context of book collections. One such function that is of great help in public service work is:

- **to *relate* subjects to other "outside" topics both (a) through formal cross-references of BT, RT, and NT relations, and (b) through displays of alphabetically adjacent subjects whose connections to each other are not caught by formal cross-references.**

I have already exemplified point (a) previously. Point (b) may not be as familiar, so let me give an example of it: in LCSH **Monasteries** is linked to the *narrower term* **Monasteries, Coptic** *not by an NT reference*, but *simply by its alphabetical proximity.* **Monasteries** is similarly linked to the cross-reference **Monasteries, Cistercian USE Cistercian Monasteries.** And the alphabetical proximity of **Monasticism and religious orders** leads to *its* NT cross-references to **Child oblates, Clerks regular, Contemplative orders** and a host of other headings otherwise scattered imperceptibly throughout the alphabet. There are whole columns of headings related to **Monasteries**–which will lead researchers in many directions–that are not linked to each other by cross-references; but they are linked nonetheless by this other mechanism. A very brief display of only some of these contiguous related headings includes the following:

Monasteries
    (linked to BX2460-BX2749 Catholic Church and NA4850 Architecture)
Monasteries, Armenian
Monasteries, Buddhist
Monasteries, Hindu
    (linked to BL1243.72-BL1243.78)
Monasteries, Jaina
    (linked to BL1378)
Monasteries, Syrian Orthodox
Monasteries and state
Monasteries in art
Monastery gardens
Monastic and religious life
    (linked to BX2435)
    BT Spiritual life–Christianity
    RT Vows
    SA *subdivision* Spiritual life *under names of individual religious orders*
    NT Celibacy–Christianity
        Eremetic life
        Evangelical counsels
        Retreats for members of religious orders
        Spiritual direction
        Superiors, religious
    –History–Early Church, ca. 30-600
        (linked to BX2465)
Monastic and religious life (Buddhism)
Monastic and religious life (Hinduism)
    (linked to BL12266.85)
Monastic and religious life (Zen Buddhism)
Monastic and religious life in art

Monastic and religious life in literature
Monastic and religious life of women
  (linked to BX4210-BX4216)
  –Psychology
    (linked to BV4205)
Monastic guest houses
  USE Monasteries–Guest accomodations
Monastic libraries
  (linked to Z675.M7)
Monastic profession
  USE Profession (in religious orders, congregations, etc.)
Monasticism and religious orders
  (BX385 Greek church)
  (BX580-BX583 Russian church)
  (BX2410-BX4560 Catholic church)

All of these displayed relationships and linkages–and scores more not listed here–would be lost without both precoordination and alphabetically-adjacent listing. Without the perceptible contiguity of **Monasteries** to these other headings, all of these paths to related knowledge records could never be noticed by researchers. (Nor, again, are they captured by the cross-referencing system of BT, RT, and NT.)

My experience in standing over researchers' shoulders and explaining LCSH to them is that very few people realize the extent, variety, and specificity of the terms available to them, without some such display enabling them to recognize the related terms they could never specify in advance via Boolean combinations. Researchers very much appreciate having these option-displays pointed out to them–they cannot think of them on their own.

Again, almost all of the alphabetically-adjacent related or narrower terms are themselves precoordinated phrases. Both their contiguity and their very existence, however, would vanish in a faceted LCSH system shackled exclusively to a postcoordinate search capability.

The Continuing Need for Reference Assistance, Over and Above Catalog Improvements, in the Total System

Doing research in large libraries is seldom "transparent" to users, even to those who limit themselves to the library's catalog; some instruction, either beforehand or at the point of use, is usually required. Without such guidance from reference librarians researchers routinely miss most of "what the library has"–let alone "what the Web has"–without realizing they've missed anything. Again, it's like Piper Cub pilots trying to fly C5-A transports; without some additional instruction, all they will be able to do on their own is taxi the larger plane along the ground–they won't be able to really exploit its heavy lifting capabilities. (This is why I say catalogs alone cannot bear the burden of doing "everything" by themselves; in the operation of the total system, reference librarians are just as integral as catalogs and catalogers if the heavy lifting capability is not to be abandoned. And our culture requires the continuance of that capability.)

I think the present Conference would not be prudent if it were to assume, without any argument, that reducing the display potential of LCSH headings, dumbing down the complexity of the strings themselves, abandoning displays of their cross-reference connectsions, and severing their links to LCC, is the way to enable people to do *better* research: to exploit that "heavy-lifting" capacity needed in large libraries. We should indeed be aiming at that goal of promoting better research; but we should also realize that its accomplishment will necessarily entail many more factors than improving library catalogs alone. One such factor is providing reference help.

LCSH Unlike Other Thesauri

An additional fact that tends to be overlooked by anyone who would reduce LCSH to the shackles of faceted thesauri is that other controlled vocabularies deal almost exclusively with the literature of one topic area; LCSH, on the other hand, must deal not only with all possible subjects of knowledge–not just information–records, but with the endless relationships between and among them, in ways that elude simple Boolean and proximity searching. (Look again at the cross-reference, and alphabetical-adjacency, examples of **Women** and **Afro-Americans**.) Other thesauri, too, (save for the *Sears List* and its links to DDC) do not have to serve as subject indexes to classification systems for shelving full-texts in arrays that allow them to be quickly browsed down to the page and paragraph level.

Significant Differences Between OPAC Cataloging and Web Metadata: The Importance of Browse Displays of Precoordinated Strings

Yet another function of *cataloging* that shows up so often in the public service context is:

- **to relate the various aspects "within" one and the same subject to each other through browse displays of subdivisions within online library catalogs.**

Most of these subdivisions are "free floaters" and, like facets in other controlled vocabularies, are not displayed as linked to their parent term within the thesaurus itself. *The needed display, however, is accomplished elsewhere, within the catalog rather than within the thesaurus.*

In other words, to point out that many LCSH strings (i.e., those with free-floating subdivisions not recorded in the thesaurus) are not displayed precoordinately within the thesaurus itself is not an argument on behalf of saying, therefore, that *all* secondary terms in any string can be treated as "free floating." This is literally a non-sequitur. Those free-floating subdivisions that are not displayed precoordinately in the list have two important characteristics: **a)** their ordering in relation to their heading is not needed to determine meaning, cross-referencing, or linkage to LCC; and **b)** their ordering in relation to their heading is indeed displayed precoordinately elsewhere, within OPAC browse displays. Even "faceted" free-floating subdivisions require precoordinated browse displays in OPACs–for without such recognition arrays, most researchers would never think of their existence in Boolean combinations. OPAC browse displays of contiguous subdivisions provide a structure that shows the extent of the subject's aspects–a structure that could never be guessed at by naive researchers entering unfamiliar subject territories.

For example, I have helped many readers who were interested in researching particular countries. One asked for help on the history of Yugoslavia. On his own he had tried a keyword search, but the Boolean combination he'd done of **Yugoslavia** AND **History** had overwhelmed him (and the computer system itself) with more than 10,000 records. So I showed him how to do a browse search that would bring up a full array of subdivisions under "Yugoslavia"; and of course this kind of display alerts the researcher to much more than the one subdivision "History." It also displays options such as:

Yugoslavia–Antiquities
Yugoslavia–Boundaries
Yugoslavia–Civilization
Yugoslavia–Description and travel
Yugoslavia–Economic conditions
Yugoslavia–Ethnic relations
Yugoslavia–Foreign relations
Yugoslavia–Intellectual life
Yugoslavia–Politics and government
Yugoslavia–Rural conditions
Yugoslavia–Social life and customs

I didn't stay to watch which aspects he chose; I just showed him how to scroll through the array. (He did get very excited when he saw "Antiquities" as an option, however.) The point is that all of these options might well be of interest to an historian of this (or any other country); but most researchers would never become aware of the *range of options* they have in researching such a topic without such a display. Further, several of these subdivisions are free-floaters not recorded in the LCSH thesaurus itself; but they do show up in the OPAC browse display. *All of these relevant paths would be lost*–and in fact *were* lost–in the reader's search for **Yugoslavia** AND **History** in a postcoordinate Boolean combination of separated facets.

Precoordinated Subdivision Strings Do Much More Than Just "Break Up Large Files"

The virtue of such precoordinate displays is not merely that they "break up large files" but that *they alert readers to whole areas of options relevant to their interests that they could not specify in advance.* Granted, if their only function were to "break up large files," then such break-ups could be done through postcoordination. But, contrary to the beliefs of some catalogers who evidently do not work with the public, this is *by no means the only function* of precoordinated subdivisions; and the "little error" of holding a mistaken assumption here will lead to "very serious consequences" for researchers who want not just "something" on their topic, but a structured overview of their research options. (I may not be articulating this very well, but the difference here is at least *like* the difference between information and knowledge–the levels relationship and interconnectivity are simply not the same.)

The Need for Recognition Capability When Prior Specification Cannot Work

One more (brief) example: I once helped a Classics professor who wanted to know how the Greeks would have transcribed animal sounds (e.g., quack, oink, meow). He was already familiar with the frogs' croaking recorded in Aristophanes' *The Frogs*; but he was interested in other animal sounds. The LCSH term **Animal sounds** looked promising, but wasn't; it just didn't work. (It did work, however, in the printed *Social Sciences and Humanities Index* to turn up an article on "Suetonius' Catalog of Animal Sounds"–a Latin list, apparently, that the professor said he would also pursue.) So I thought we might browse through the subdivisions under **Greek language** to see what might turn up. What did turn up was **Greek language–Onomatopoeic words**, which led to a dictionary that included animal sounds. (I don't read Greek myself, but the professor told me he was satisfied with the book.)

Now of course it could be said that a postcoordinate combination of **Greek language** AND **Onomatopoei?** would turn up the same result; and that would be a true statement. But it would also entirely miss the point: Who would ever think in advance to use "Onomatopoei?" as one of the elements *in* the combination? (Similarly, who would think beforehand of all the differently-phrased options under "Yugoslavia"?) The major virtue of precoodinated displays of subject strings is that *they bring to our attention options that we could never specify in advance.* And the larger the file that is being indexed/cataloged, the more necessary are such aids if the resultant retrieval is to be anything more than fragmentary and orphaned from relatives. Again, it's roughly the difference between finding *information* about a few isolated options you can specify, vs. gaining a *knowledgeable overview*–a map that shows both the existence and the relationships–of all of your options within the catalog. (Writers who rhapsodize about the wonderful ways of searching brought about by computers seldom mention how much *more* powerful the computer searches themselves become when they enable readers to see precoordinated strings in browse displays–displays that enlist the tremendous power of simple *recognition*.)

Catalogs Cannot Do Everything That Needs To Be Done

Let me also add that in the "Yugoslavia" case I also put under the reader's nose the wonderful current article on the country in *Europa Yearbook*, and the *Yugoslavia: A Country Study* (1992) volume from the old area handbook series. And I let him know that we could also easily find a variety of other concise overview articles from scores of specialized encyclopedias by using the *First Stop* and *Subject Encyclopedias* indexes (neither of which is computerized). There is no way on earth this man would have found these overview starting-points on his own by searching the library's catalog, especially with **Yugoslavia AND History**. Even if he'd seen the record for the area handbook volume–which does not have the word "History" anywhere on it–its special significance as a starting-point would not have leapt out at him.

Once again: the catalog alone simply cannot do *everything* that needs to be done for researchers; and this Conference should not be assuming that it needs to take on that function. F. W. Lancaster, in his "Second Thoughts on the Paperless Society" article[13], makes some cogent observations:

> The [library/information science] profession has greatly exaggerated the benefits of technology, especially in the area of subject access. Putting electronic databases in the hands of library users does not necessarily mean that they can be used effectively. . . . Merging several catalogs into one creates much larger databases that are even less useful for subject access than their components. . . . Unfortunately, the majority of librarians seem to assume that more access means better access. This is not necessarily true. For 30 years, studies have consistently shown that information services users really want access to the best information. They want tools or people capable of separating the wheat from the chaff. They want quality filtering.
>
> The profession seems to have lost sight of this. How else can one explain why so many librarians are head over heels in love with the Internet, a monster lacking a minimum of control of content? . . .
>
> The service ideal still exists to some extent in public libraries and school libraries. However, the more specialized the library becomes in the academic world, encouraging remote use, the more dehumanized it becomes. [The more, too, it trades away orientation to knowledge for access to information–TM.] The closer the professional is to the public, the more the service ideal survives and will continue to do so.

The Strengths and Weaknesses of Catalogs

In providing subject access, if there is one thing that library catalogs are good for it is in providing *overviews of search options* through displays of precoordinated subject headings and subject-subdivision strings. (Of course catalogs do other things, too.)

If there is one thing that they are notoriously bad for, it is in separating the wheat from the chaff–of pointing out the *best* individual sources from the many arrays and categories of options. (The fact that they point to professionally selected collections, however, puts them in marked contrast to Web search engines.) Library catalogs are also incompetent to lead readers to the best databases for journal articles among the hundreds available, let alone to the best articles themselves[14]; or to starting-point/orientation articles in the thousands of specialized encyclopedias that are not available online. Catalogs also have weaknesses in bringing to readers' attention government documents, microform research collections, and special collections. There are *other ways* to get into such things, however, as any good reference librarian knows. We don't need library catalogs to take on all of these functions–to "seamlessly" integrate "vast resources" all in one overwhelming source. The catalog is *one* necessary avenue of access to *some* necessary records; to overburden it with too many functions would be to kill a goose that lays golden eggs, and to undercut its ability to turn up *books* in particular. (Better home pages or portals that lead to the catalog *in relation to* other sources, could help here; but the catalog *itself* cannot lead seamlessly to *all* necessary sources–nor, for that matter, can even the best home pages or portals.)

The importance of seams

The larger point here is that visible "seams" among resources are in fact necessary for researchers. When a portal screen tells a researcher, in effect, to click *here* for access to books, *here* for journal articles, *here* for dissertations, *here* for Web sites, and *here* for newspaper articles, and so on–when it *shows the seams*, in other words, it thereby provides a structured overview of options that would otherwise be imperceptible. One of the greatest frustrations researchers have is that of not knowing "where they are"–of not knowing the extent of the results they initially retrieve, and whether they are looking at "everything." Seams between and among research options help readers to recognize a variety of paths that they can follow if their initial results are inadequate. Seams serve as perceptible *boundaries that provide points of reference*; without such boundaries, readers get "lost at sea" and don't know where they are in relation to anything else: they can't perceive either the extent of what they have, or of what they don't have.

Automated Collocation?

No other source–not *Books in Print* (with its inadequately subdivided subject headings), not Amazon.com, not Google–is as good at finding *books* by subject as a good library catalog. Automated means of subject collocation are no substitute for good cataloging. In Amazon.com, for example, the record for my own book, *The Oxford Guide to Library Research*, adds the following helpful notice:

> Customers who bought titles by Thomas Mann also bought titles by these authors:

- Franz Kafka
- J. K. Rowling
- Herman Hesse
- Andre Gide
- Feodor Dostoevsky

Much as I would wish to offer this as an example of the extraordinary insight, accuracy, and trustworthiness of the collocation software, I fear that more objective observers may reasonably conclude that a Large Mistake Has Been Made.

Catalogers Reading from Different Page?

(This is just an impression, but perhaps it's relevant: Much of the library world is trying to find reasons to induce people to continue coming inside the library's walls–*and pay their tax monies for supporting those walls and the nondigitized collections within them*–instead of just searching the Internet from their homes, schools, or offices. The cataloging wing of our profession, however, sometimes seems determined to create a product that will seamlessly cover "everything"–especially the Internet, which does not require entry within library walls–and do it in such a way that the catalog product itself can be tapped into by anyone, from anywhere, at any time. [The title of a recent conference of the New England Technical Services Librarians was "User Oriented Technical Services: All Things to All People."] It would help if catalogers would start thinking *outside the box of the Internet alone*, and realize how many important things–especially copyrighted books–are not *in* that Internet box, but still need good localized access and arrangement mechanisms. In other words, it might help to preserve libraries-as-places if catalogers were reading from the same page as the rest of us.)

Significant Differences Between OPAC Cataloging and Web Metadata: LCSH's Inextricable Links to LCC

Yet another function of cataloging–unlike metadata–is, again:

- **to serve as the functional index to the Library of Congress Classification scheme (LCC) in the classified bookstacks.**

It is through the subject headings in a library catalog, and their links to records with different class numbers, that researchers are enabled most efficiently to discover which areas of the stacks they need to go to (and which to avoid) for in-depth browsing of full texts of books on particular subjects. Without this linkage, which appears within catalogs themselves more than in the LCSH list (although the linkage is there, too, to a lesser extent), the exploitation of classified bookstacks would be greatly undercut, as it would not be easily determinable which stack areas cover which subjects. (Readers use library catalogs to index the bookstacks–there is no way they are going to endure catalogers' indexes to LCC.)

The Continuing Need for Subject-Classified Bookstacks

The advantages of classified bookstacks are that they allow in-depth subject searching of *full-texts, not just catalog records*–i.e, readers can quickly scan whole groups of related texts *right next to each other*, not just for tables of contents, but also for running heads, illustrations, maps, charts, portraits, diagrams, statistical tables, highlighted boxes, typographical and color variations for emphasis, marginalia, footnotes, bibliographies, and indexes at the backs of books–none of which is digitized on catalog records. (Nor are the vast majority of the hundred thousand+ copyrighted books published each year making the "shift to digital" forms that Mr. Dillon apparently assumes; significantly, Mr. Dillon's own book itself has not made the shift.[15])

LCSH Must Function in Both Book and Web Contexts

The future of LCSH, in other words, must be planned with the maintenance of *this* context in mind, just as much as a Web context. Research libraries–unlike many special libraries–must continue to operate in both the contexts of online resources and print collections. It is not a matter of one context *rather than* the other, or one *superseding* the other, or one *shifting* to the other, or one *evolving into* the other. The requirements of discovering the knowledge contents of large book collections are not the same as those of searching the Web for unintegrated and unrelated information (which is, and probably will continue to be, the Web's primary–not only, but primary–function).

There are thus two contexts for the future use of LCSH, and the *book-collection context will not go away*. Nor can it be forced onto a Procrustean bed of postcoordinate search mechanisms more appropriate to the Web context without decimating the efficiency and "heavy lift" capacity of catalogs in providing subject access to large book collections.

This is, then, a real problem with some of the papers on the Bicentennial Conference Web site: They look at the future of LCSH *exclusively in the one context of Web resources*. (Pardon my redundancy; the point needs emphasis.) The "little error" of such a blinkered initial assumption will lead to "very serious consequences" for historians, biographers, literary scholars, and researchers in general who will, and often must, continue to use the vast stores of knowledge records, both retrospective and current, that simply are not and never will be digitized on the Web.

Missing Stakeholders

By the way, where are the representatives of stakeholders such as the American Historical Association, or the Organization of American Historians, the American Association of University Professors, or the associations of the other scholarly interests? If, by chance, the result of our Conference is to radically change the way *books* are given subject cataloging–so that future headings no longer show up in browse displays related to existing headings; or so that the library catalog no longer functions as an index to the classified bookstacks–then shouldn't groups of professional academics who *depend on the book collections of research libraries* have a seat at the table? Surely we are not going to unilaterally declare that they will no longer need efficient subject access to large book collections in the future! How would *The Chronicle of Higher Education*, *Lingua Franca* and Nicholson Baker report such chutzpah?

Summary of Differences

It is highly unlikely that anyone will ever consider *library catalogs* as their first choice of entry into the Web–not at least, until library catalogs cover the *billion+* records indexed by Google et al. There are about 95,000 records in the RLG Union Catalog that point to digital resources (that is, having 856 fields)[16]; and we all hope this Conference will find ways to expedite the inclusion of still more such resources into library catalogs. But if we disregard, in our initial assumptions, the very features that make library catalogs such useful guides to substantive knowledge records then we will have done more damage than good to higher education in this country. Library catalogs and LCSH, unlike Web search engines with faceted metadata, have these features:

- They are tied, to begin with, to substantive, professionally selected records–books–that are proven media for conveying knowledge, not just information, and that can be economically preserved for centuries;

- They relate and link different subjects to each other in cross-disciplinary ways;

- They spell out the many unforseen aspects that lie (otherwise indistinguishably and unnoticeably) within any one subject field;

- They allow researchers to recognize relevant topics and relationships that they could never specify in advance;

- They guide researchers most efficiently to one or more areas of the bookstacks (rather than others), where so many of the substantive and non-digital knowledge records reside for quick browsing down to the page and paragraph levels.

The latter four functions are highly dependent on precoordination.

Blurred Distinctions

Two very important distinctions seem to be getting blurred in some of the papers before this conference:

1) Should the future of LCSH be considered primarily in terms of Web-type search softwares that do not allow precoordinate displays of subject strings–i.e., should it be our goal to change OPAC softwares themselves to be more like Google?

2) When we talk about extending LCSH to cover Web resources, do we mean:

(a) "covering" Web resources by creating surrogate catalog records for them, just as we do for books, which will show up "in the catalog"–i.e., within OPAC browse displays of precoordinated strings (as in the **Women–Services for** example at the beginning of this paper) in relation to the other surrogates already in the catalog?

Or do we mean:

(b) somehow adding LCSH elements directly to the headers of the actual Web records ("applications of metadata") out in the Internet–i.e., to headers residing within the Web sites themselves, not to surrogates merely pointing to them from their residence in the OPAC?

Intellectual Property Issues

Regarding (1): Given the billion+ Web sites that already exist, and the Web's rate of growth, isn't it just common sense to regard the application by catalogers of LCSH metadata elements to the headers of Web records, directly, to be a hopelessly Sisyphean task? Isn't it common sense also, to begin with, to recognize that *we do not have the authority to tamper directly with the intellectual property of billions of Webmasters by obtruding our presence into their sites*? We can do anything we want with *surrogate catalog records that we create in our own OPACs*; but we simply have no right to tamper directly with the metadata on headers within Web records themselves.

Perhaps, then, we can suggest improvements, not to the countless Webmasters' sites themselves, but to the commercial engines like Google and NorthernLight, et al., which *index* those sites. That is, perhaps we can recommend ways in which their weighting and ranking softwares can be tied to authority lists, in order to map words in retrieval results to faceted LCSH elements, which would provide some additional measure of control to the keyword-weighting process. (Precoordinated strings would be out of the question in this context–no machine could assign them automatically.)

I have no objection whatever to our making suggestions to the search engines that we do not control ourselves. But in the blur of these distinctions, I would emphatically remind everyone, again, that intellectual property rights are involved: *librarians do not and cannot control these*

commercial Google-type indexing enterprises any more than we can control the Webmaster-created headers of the Web sites they index.

Merging OPAC Searching with Internet Searching?

The only things we can *control* are the things we create ourselves. This means *library catalogs*, not Google or HotBot or their commercial cousins. If we confine ourselves to examining the future of library catalogs–the only things we can control–then we have different options:

Option A: We can attempt to merge the searching of library OPACs with the searching of Internet sites through software changes. This merging could theoretically be done "from the outside in," or "from the inside out":

A.1. "From the outside in." We could abandon our existing OPAC softwares for searching bodies of catalog records separated from the Web. By merging our catalogs into the Web we could open their full contents directly to Web search engines such as Google or Yahoo. We could simply piggyback on these existing services already known to, and widely used by, researchers. A Google search of the future, then, would seamlessly turn up surrogate catalog records for books, created by librarians, in the same operations that retrieve Web sites created by others. We could continue to assign LCSH elements that would serve as metadata elements searchable by Google type engines rather than by segregated OPAC softwares. Since Web engines cannot show precoordinated strings in browse displays, we should simply abandon precoordination in LCSH.

A.2. " From the inside out." We could radically change our own library catalogs so that they, *like* Google, try automatically to index not just local collections-within-walls but the entire Web, via spiders, crawlers, harvesters, and term-weighters *of our own creation*. Unlike Google, however, *our* automated indexes could add faceted LCSH elements through softwares that would map weighted keywords to controlled LCSH elements, whether or not these elements appear in the headers or bodies of the indexed sites Web sites themselves that exist beyond our own catalog records. While, for intellectual property reasons, we could not force LCSH elements into the headers of Web sites created by others, our software could add them to the displayed *results* of weighted keyword searches, to provide additional elements of control not otherwise present. This option, too, would necessarily abandon the display of precoordinated strings of LCSH terms, because no mapping software could possibly create proper strings, or displayed linkages among them, simply on the basis of weighted keywords.

If we go in the direction of **Option A,** in either of its variants, we would effectively have to merge catalog records for books–which we would continue to create–into the same "pool" as the Web environment that we seek to catalog, and which already exists outside our present catalogs. The major difference lies in whether we search the records by existing external softwares (from the outside in) or through new softwares of our own devising (from the inside out). In neither case would there be any point to continuing precoordination in LCSH, since neither option would be capable of showing subject heading strings in browse displays.

Book Records Buried in Chaff, Loss of Connection Between LCSH and LCC

Before considering option (B) for the future of library catalogs, let me say why I think option (A) is unworkable. First and foremost, even if faceted LCSH terms were somehow mapped automatically to all Web sites *and* added manually by catalogers to individual book sites, the book records would become so buried within the overwhelming chaff of the Web that researchers would no longer be able even to identify the ones most relevant to their topics. Nor would researchers be able to view such records for books in relationship to other book records–or, for that matter, identify books in relation to the most relevant Web sites.

There would just be too much chaff; and the assignment of faceted LCSH elements would simply not be enough to control retrieval in any way noticeably better than what Google does.

Such a Web-search library catalog would utterly sever the existing network of strong connections *from* book records cataloged with precoordinated LCSH elements *to* particular areas of their local classified book collections. This would effectively vitiate the possibility of scholars efficiently browsing classified book collections locally.

I think we may reasonably conclude that future catalogs should not, like Google or Hotbot, try to swallow the whole Internet or to merge into it; they will maintain their utility only by indexing highly-selected portions of the Web, and in a way that does not overwhelm researchers with unwanted chaff.

Expanding the Range of Free-Floating Form Subdivisions to Include Web Sites

A second option for the future of library catalogs would be:

Option B: We could continue to use the software of existing library catalogs that show browse displays of precoordinated LCSH headings, but expand the range of (free-floating) subdivisions to include form subdivisions for Web sites. Let me repeat here the example given earlier:

Women–Services for
Women–Services for–Bolivia–Directories
Women–Services for–Caribbean area–Case studies
Women–Service for–Ethiopia–Congresses

Women–Services for–Germany–History
Women–Services for–Michigan–Evaluation
Women–Services for–New Zealand–Bibliography
Women–Services for–North Carolina–Finance
Women–Services for–Study and teaching–United States
Women–Services for–Study and teaching–United States–Web sites (.edu)
Women–Services for–United States–Directories
Women–Services for–United States–Web sites (.com)
Women–Services for–United States–Web sites (.edu)
Women–Services for–United States–Web sites (.edu)–Data archives
[Again, "Data archives" may not be an appropriate subdivision for this particular subject; I offer it here just as a pattern example.]
Women–Services for–United States–Web sites (.edu)–Discussion lists
Women–Services for–United States–Web sites (.edu)–Portals
[Again, "–Site directories" might be an alternative, in which case a cross-reference is needed: Site directories USE Portals]
Women–Services for–United States–Web sites (.gov)
Women–Services for–United States–Web sites (.org)
Women–Services for–Wisconsin–Periodicals
Women–Services for–Zambia–Directories

Of course, live links would be provided from the catalog surrogates to the actual Web sites, insofar as licensing agreements allow.

Precoordinated displays like this in OPACs would (1) separate the substantive Web sites from the clutter of chaff turned up by Web search engines, and (2) show them in relationship to scholarly book records–an ideal outcome. We would be using precisely the strengths of the catalog in its unique display potential, as well as in its selectivity, to overcome the weaknesses of the Web. These goals ought to be at least part of what we are aiming for.

The Large Question

But we need to do more than just this. **The larger question before this Conference, I think, is this: How can we (a) simultaneously get LCSH into both metadata fields of Web records created by other people *and* into the OPACs that we create ourselves; and (b) do it in a way that will simultaneously exploit the strengths of both the flexible postcoordinating software of existing Web search engines *and* the powerful browse screen capabilities of OPACs?** This would be Option C, to which I shall return.

Is Loss of Precoordination Really Logical?

As a prelude to **Option C**, however, I must first comment directly on Lois Mai Chan's paper.[17] When Ms. Chan asks the question "What direction and steps need to be taken for LCSH to overcome these limitations and remain useful in its traditional roles as well as to accomodate other uses?" she specifically includes "systems with index browsing capability" among the "limitations" that must be "overcome. " She reports, further, on one of her current projects:

> Using LCSH as the source vocabulary, FAST (Faceted Application of Subject Terminology), a current OCLC research project, explores the possibility and feasibility of a postcoordinate approach by separating time, space, and form data from the subject heading string (Chan et al. in press).

She also comments, a paragraph later:

> Considering the gradual steps the Library of Congress has taken over the years, even a person not familiar with the history of LCSH must conclude logically that LCSH is heading in the direction of becoming a fully faceted vocabulary. It is not there yet; but, with further effort . . .

The phrase "not there yet" obviously implies an acceptance, and recommendation, of what seems to be a "logically" inevitable transformation of LCSH into a system of fully faceted elements (which can only be contrasted with a system of precoordinated strings). These comments, however, need to be placed in the context of another very recent paper by Ms. Chan, appearing in *Cataloging & Classification Quarterly,*[18] in which she writes:

> Within the OPAC environment, where trained personnel is available for the creation and maintenance of complex subject heading strings and the online system is capable of handling such, the current rules and policies for complex syntax can continue to function.

Amen. This point, I think, needs much greater emphasis than it receives in Ms. Chan's paper before the present Conference. The **Option C** that I will propose is one that I think (hope?) we can agree on; but here is the key point: we must consider the future of LCSH, as I have argued above, in *two* continuing environments that are *very different* from each other: one, the OPAC/book-collection environment, and the other, the Web/networked environment. And *because* the book collection environment will not transform, merge, or evolve into the Web/networked environment but will always remain distinct from it, I maintain that we need a future LCSH that does not lose the many existing strengths of precoordinate displays. This is the crucial difference: one environment supports the display of precoordinate LCSH strings and the other simply does not.

What I am afraid of is that Ms. Chan's conference paper readily lends itself to misinterpretation, because while it does indeed recognize (some) important distinctions between the two environments, its portrayal of the "logical" future of LCSH in the Web/networked environment *silently entails its loss of precoordination in the OPAC/book collection environment*–unless Ms. Chan advocates that two different LCSH systems be maintained in the future for the two different environments. She is silent on this; but I suspect she (and everyone else) would regard the maintenance of two different LCSH systems to be economically as well as intellectually unsupportable.

What Would Be Lost

The theoretically extrapolated loss of precoordination, however, is neither logical, nor necessary, nor inevitable, nor desirable:

- It is not logical to abandon precoordination when the *very meaning* of so many LCSH terms is dependent on the word-order of their phrasing, in ways that cannot be recaptured by postcoordinate Boolean combinations *or* by word-proximity searches that drop out relational prepositions as stopwords.

- It is not logical to abandon precoordination when to do so would uproot tens of thousands of LCSH strings from a *vast web of specific linkages to LCC*–i.e., changes in the word order of the subject strings *also* changes the classification areas to which they point.

- It is not logical to abandon precoordination when browse displays of subject-string phrases *enable researchers simply to recognize* whole ranges of options that they could never specify in advance through postcoordinate combinations (e.g., **Yugoslavia–Antiquities** rather than just **Yugoslavia** AND **History**; **Afro-American whalers** rather than just **Afro-Americans** AND **History**; **Greek language–Onomatopoeic words** rather than just **Animal sounds**). The larger the file, the more researchers are dependent on *recognition* of options that they cannot articulate beforehand.

- It is not logical to abandon precoordination when the existence of the *vast cross-reference structure* between and among headings is so heavily dependent on the retention of ordered strings (e.g, **Women-Psychology** NT **Leadership in women**; **Afro-Americans–Education** NT **Segregation in higher education–United States**).

- It is not logical to abandon precoordination when the *relationships of alphabetically-adjacent headings* within the thesaurus would be entirely lost without it (e.g., **Monasteries** is linked to scores of precoordinated neighbor headings such as **Monasteries and state** and **Monastic and religious life of women** simply by their displayed contiguity rather than by any formal cross-references).

- It is not logical to abandon precoordination when LCSH, unlike any other thesaurus, must simultaneously cover *all subject areas–not just one, as other thesauri do–and show relationships among them* that readers could not specify in advance.

Nine years ago Ms. Chan read a paper to the Airlee House Conference, similarly calling for less precoordination and greater use of postcoordinate combinations of individual, faceted elements in LCSH. The members of that conference listened respectfully, but then ignored the substance of the paper–i.e., the course of the subsequent discussion immediately became, effectively, not "Should there be less precoordination?" but rather "*Given the need to retain precoordination* [for the above reasons], what should be *the order of the string elements*?" Subsequent improvements in search software–as in Google, Hotbot, et al., which did not exist at the time–have not invalidated any of the above reasons for retaining precoordination in LCSH.

A theoretically-extrapolated projection of greater postcoordination of individual facets simply ignores the reality of the many functions LCSH already serves in the real world of real library collections; and these continuing (and growing) functions are just as much a part of its history as is the trend to break phrase headings into subdivided (*but still precoordinated*) strings in browse displays. The real world of practice and function puts *real and definite limits* on the "direction" of LCSH toward "becoming a fully faceted vocabulary." None of these realities is given anything more than passing mention–most are not even mentioned at all–in Ms. Chan's current paper. This will never do; such "errors in the beginning lead to serious consequences in the end."

While greater facetization–if there is such a word–of LCSH may indeed be a desirable goal *in a Web environment such as* **Option A** *above*, in which we abandon our current OPAC softwares, I think we need to question whether **Option A** is even possible, let alone desirable, to begin with. One crucial point is that Ms. Chan simply does not consider the question of intellectual property: Can librarians add LCSH elements to headers of countless Web sites whose Webmasters have no obligation whatever to pay any attention to what librarians want? Answer: No, we cannot. If, then, librarians cannot obtrude our terms into other people's intellectual property sites, what chance do we have of getting independent Webmasters to voluntarily start using LCSH elements in their headers? And what will the results of LCSH, either faceted or precoordinated, applied by rank amateurs be like? Will it sustain the "heavy lifting" capacity that our large research libraries–and our nation's intellectual culture itself–require? The results of utterly fragmented LCSH elements applied as metadata to Web headers by amateurs, I suspect, would hardly bear any relation to what professionals usually think of as "vocabulary control." (And how do we prevent Webmasters of porno sites from having a field day with their voluntary use of LCSH's **Women** terms in their headers?)

The larger question here, of course, is this: Should our profession consider the primary future use of LCSH to be by Webmasters *over whom we have no control*? I think not.

Getting Librarian-Created LCSH Elements Into the Headers of Web Sites

But I also think there *is* a way that we can get professional-librarian-assigned LCSH elements into the headers maintained by independent Webmasters. This is a proposal is similar to what Regina

Reynolds recommends in her paper, and in line with what Priscilla Caplan recommends when she calls us to "work proactively with publishers."[19] It is:

> **Option C: an Online Cataloging In Publication (OCIP) program that mirrors our current CIP program for printed books.** With such librarian-created metadata added to the Web sites of *quality-screened participants*, we would have the best of both environments: We could continue to assign LCSH in traditional precoordinated strings on the *surrogate records that we create*–but these records would then appear *in both environments*: directly within the program's Web records as metadata in their headers *and* simultaneously in OPACs as catalog records.
>
> In the Web environment, as metadata, even if the LCSH elements are assigned as strings, their individual words or facets could still be searched postcoordinately by *existing* services such as Google and NorthernLight, without our having to overhaul our own expensive catalog softwares.
>
> In the OPAC environment, in contrast, the same LCSH elements could still be searched in their precoordinated forms in catalog browse displays–as well as postcoordinately. Their preocoodinated display, as with the **Women–Services for** example above, would *relate* the quality-selected Web sites to existing and future *book* records, as well as to other quality-selected Web sites–and also do it is a way that does not undercut the widespread linkages of LCSH to LCC in the classified bookstacks, nor undercut the cross-reference structure, undercut users' recognition capabilities, etc., etc.
>
> I do not mean to suggest that library catalogers should create catalog records *only* for Webmasters who sign up for the OCIP program. Far from it. Library catalogers should be free to create surrogate catalog records in their OPACs that point to any Web site at all worthy of being brought to researchers' attention. And there is nothing in this OCIP proposal to prevent this. The *extra* advantage of an OCIP program, however, is that the cataloging data created for participants in the program would *also* become searchable as metadata in the participants Web sites–i.e., accessible not just on catalog surrogates through library OPACs but also within metadata fields accessible via Google and HotBot and all the other engines.
>
> This proposal also has the advantage of saving us the expense of radically redesigning the expensive search softwares of our existing OPACs. And it includes all of the strengths of **Option B** while also averting the intellectual property problems, and those of overwhelming chaff, in **Option A**.
>
> Yet another likely advantage: if the existence of the OCIP program were made known as widely among Webmasters–especially corporate bodies–as CIP is among

> publishers, then the Webmasters of high-quality sites will probably start trying to bring their sites to our attention, on their own initiative. Just as CIP records make books more attractive to libraries, OCIP records would make Web sites similarly attractive. To get into the program, however, Webmasters would have to document both the quality and the likely longevity of their sites for us. That means librarians wouldn't have to spend endless time surfing around, *looking* for the best sites. Their producers would strive to bring them to our attention.

Of course there is a larger managerial/administrative problem to be worked out: Should the Library of Congress be the *only* library responsible for creating OCIP records, as with CIP records? I think this is inadvisable. Given the sheer size of the Web, and the number of possible applicants for participation in the program, the work would have to be divvied up. I think that can be managed. (Perhaps division by States, with first priority given within them to local .edu domain sites? [LC could concentrate on federal .gov sites.] A State-run OCIP program, administered through both State libraries and State University libraries, might also enable us to get a handle on how to divvy up electronic *preservation* responsibilities. We can't even begin to preserve everything on the Web; but perhaps the sites of OCIP participants within each State would provide an initial rough focus for preservation attention? Indeed, an increased likelihood of preservation might well serve as an incentive for Webmasters to join the program.) The details are outside the scope of this paper, and probably outside my own competence to imagine.

Doesn't **Option C**, however, address many of the major problems confronting this Conference? Priscilla Kaplan says in her paper, "The most critical factor in the future of DCMES [Dublin Core Metadata Element Set] is whether a working organization can be achieved to manage the change process and to produce the documentation, support structures, and policies required by an international community of implementers holding very little in common."[20] I suspect an OCIP program–probably having to extend beyond U.S. States to foreign participants–holds the best hope of creating the *locus* sites that will be necessary to create these support structures.

The Need for Consistency and Accuracy in Subject Heading Assignment

There is one further issue that I think this Conference needs to address squarely: If we are going to use LCSH in *both* OPAC and Web environments of the future–and I heartily hope that we will–it really does make a difference that we strive for consistency and accuracy of subject-heading assignment. There isn't any "control" in "vocabulary control" to begin with if subject cataloging is relegated to low level technicians who know nothing of specific entry or cross-references. Nor can there be much control if we regard Web sites rather than books as the primary targets of our cataloging activities, for the simple reason that LCSH elements appearing in metadata fields, *if considered only as separate from OPAC displays of the same data*, do not require the many extra controls of precoordination, cross-referencing, links to LCC, or displayed alphabetical adjacency to related headings.

Are Web Sites More Important Than Books?

This, then, brings us to some of the proposals put forward by my former LC colleague Sarah Thomas, which she makes in her paper, "The Catalog as Portal to the Internet."[21] There are many, many worthwhile observations in this paper. But then it comes to:

> 1. We should decisively reduce the amount of time we devote to the cataloging of books in order to reallocate the time of our bibliographic control experts to provide access to other resources, especially Internet resources . . .

I thank Ms. Thomas for a bluntness comparable to Mr. Dillon's. It is easier to engage in healthy debate when one's assumptions are not buried as concealed propositions. The forthright message here is that books are now of less importance to our culture than are Internet sites.

I beg to differ.

In the first place, our larger culture depends on libraries and librarians to provide *free* access to *books*. The full-texts of most books are not on the Internet, and most never will be, for copyright (life of author plus seventy years) and preservation reasons alone. Those that do appear, either freely available to anyone from anywhere, or free only to users of site-licensed terminals within library walls, will not be *read* online because of their lengths, but will be printed out individually at much greater-than-present costs either to libraries or to the environment, or both.

Further, it will very soon be the case that no one--not even poor people--will be dependent on libraries or librarians for access to the freely-accessible portions of the Internet[22]; but our culture as a whole will still be very much dependent on libraries and librarians for *free* access to the scores of thousands of *books* that continue to be published every year (cf. *Bowker Annual*), as well as to the low-use texts of earlier decades and centuries.

Further, all of those home- and office-connected Internet searchers will not be dependent in any way on libraries or library catalogs for ways to search the Internet: they will have Google, Hotbot, AltaVista, NorthernLight, and a wide array of other avenues of access freely available to them. Even if librarian-created catalogs are modified to include selected high-quality Internet sites (as in Options B and C above), I think it is highly unlikely that searchers would consider them as their first or most important avenues of access to the Net, in preference to Google et al. The virtue of library catalogs will lie precisely in:

(a) pointing researchers to important resources--*books*--that cannot be found on the Net to begin with;
(b) pointing them to high-quality Net sites that will otherwise be buried in the chaff turned up by Web search engines; and

(c) in relating books and quality Web sites to each other intelligibly rather than haphazardly.

But researchers will lose out on the benefits of **(a)** and **(c)** exactly to the extent that librarians, following Ms. Thomas's advice, "decisively ... reallocate" their time and attention to **(b)**. It seems that Ms. Thomas does not consider **(a)** and **(c)** as important to begin with. As a reference librarian who must help thousands of very confused researchers every year, I beg to differ. I do consider them very important.

The additional danger of slanting library catalogs primarily to Internet sites has already been alluded to (pp. 11-12 and footnote 7): We librarians and information specialists may unintentionally wind up dumbing down our larger culture if we give the primacy of our attention to a resource that is itself slanted to shorter (rather than longer) texts, visual images, audio resources, and graphical displays over textual explanations–i.e., to a medium that much more readily conveys data and information than knowledge or understanding. Again, our larger culture does not depend on librarians or library catalogs for free access to the Internet; but it very much does depend on us for free access to the substantive *alternatives* to the Net, and for the integration of the Net *into larger webs of knowledge relationships*. These needs cannot be met under Ms. Thomas's proposal for redefining our priorities.

Accepting Copy Cataloging "with little or no modification"?

Ms. Thomas then goes on to say:

> 2. In order to reduce the time spent cataloging books, we will need to investigate and implement a combination of the following:
>
> * * *
>
> Accepting copy cataloging with little or no modification from other cataloging agencies, including vendors

Ms. Thomas's enthusiasm for accepting virtually any copy cataloging "with little or no modification" has a noteworthy history. It was she who led the Library of Congress into adopting this practice in a big way. (Even now, however, it is not easy to generalize about LC's cataloging operations; there are about three dozen cataloging teams, and they vary in the level of review that they give to copied records. Some do accept copy "with little or no modification"; some don't.)

"Only about 20% agreement among catalogers"?

Ms. Thomas, in order to embark LC on the project of accepting copy-cataloging widely, invited her friend and colleague Carol Mandel, from Columbia, to address LC's troops in a Cataloging Forum meeting on 12/9/1993. There Ms. Mandel told all of us, with Ms. Thomas's approval, that "studies" show that there is "only about 20% agreement among catalogers" concerning which subject headings should be assigned. This assertion repeated Ms. Mandel's claim in her 1991 "Cataloging Must Change!" article in *Library Journal*,[23] written with Dorothy Gregor. Because of this alleged "lack of interindexer consistency," this articles says, "Catalogers can be more accepting of variations in subject choices in member copy and need not spend undue time determining whether their analyses are consistent with LC's and with those of catalogers elsewhere." Evidently on the basis of Ms. Mandel's scholarship and sources cited , Ms. Thomas herself wrote in 1993: "Recent studies have determined that intersearcher consistency does not exist. . . . ***With this new knowledge***, administrators and catalogers are asking to what extent strict consistency of application of subject headings increases the quality of the bibliographic record for use by end users'[24] [emphasis added].

The claim that there is only 20% agreement among subject catalogers was simply accepted as "knowledge" by Ms. Thomas. LC's acceptance of cataloging copy–with subject headings largely unchecked for accuracy, completeness, or consistency–shot up from 1,800 titles in 1991 to over 45,000 in 1994, under her direction.

A few years later, having come across a number of disturbingly inaccurate records that I found too late to help a few readers who could have profited from them, I began to wonder about the basis of Ms. Thomas's faith in copy cataloging that is accepted with little or no modification. So I went back to Ms. Mandel's "Cataloging Must Change!" article to check out its footnotes.

Getting the Basic Facts Wrong

What I found, briefly, is that Ms. Mandel and co-author Ms. Gregor had their facts 180 degrees backward: the studies they rely on show that the low interindexer consistency rate of ca. 20% shows up repeatedly precisely *in the absence of* vocabulary control mechanisms.[25] This is the figure achieved by amateurs who are trying to guess which keywords should be used to index a document, usually in situations entirely lacking thesauri, cross-references, familiarity with cataloging principles (especially the convention of specific entry), and established catalogs exhibiting an established pool of vocabulary-controlled records. Subsequent studies suggest that ca. 80% consistency can be expected among professional catalogers who follow the rules.[26] One, by Elaine Svenonius and Dorothy McGarry, states: "*The price that is currently being paid for lack of subject expertise in non-LC subject cataloging is that over 50 percent of the books so cataloged [i.e., by agencies other than LC] are either missing headings or have headings that are incorrect, dated, or questionable*'[27] [emphasis added].

Result of "little or no modification" in Subject Cataloging: *Subject Guide to Books in Print* Example

What is the result, for users, of bad subject cataloging? Since Ms. Thomas herself appeals to anecdotal evidence in her own paper, I will have no qualms in using it here. I would appeal to it in any event; the importance of examples, case studies, and first-hand testimony is established in many fields, including Law, beyond our own discipline.

Let's look first at subject cataloging from a commercial source. One that is readily available in libraries throughout the country is Bowker's *Subject Guide to Books in Print* (*SGBIP*). To stay within the ballpark of the **Women** examples used elsewhere in this paper, here are five examples of the subject cataloging done by the Library of Congress and *SGBIP*:

- Title: *The Beijing Declaration and the Platform for Action: Fourth World Conference on Women, Beijing, China, 4-15 September 1995.*

  LC headings: World Conference on Women (4th : 1995 : Beijing, China)
  Women–Social conditions–Congresses
  Women's rights–International cooperation–Congresses
  Women in development–International cooperation–Congresses

  SGBIP: Women

- Title: *Women as Elders: The Feminist Politics of Aging*

  LC Headings: Aged women–Congresses
  Aged women–Religious life–Congresses

  SGBIP: Women

- Title: *Female Gangs in America: Essays on Girls, Gangs and Gender*

  LC Headings: Gangs–United States
  Female juvenile delinquents–United States
  Female offenders–United States

  SGBIP: Gangs

- Title: *The Women, Gender and Development Reader*

  LC Headings: Women in development
  Women–Social conditions
  Women–Economic conditions
  Women–Developing countries

SGBIP: Women

- Title: Women Overseas: Memoirs of the Canadian Red Cross Group

LC Headings: Canadian Red Cross Society–Biography
World War, 1939-1945–War work–Red Cross
Korean War, 1950-1953–Participation, female
World War, 1939-1945–Personal narratives, Canadian
World War, 1939-1945–Participation, female
Korean War, 1950-1953–Personal narratives, Canadian
Nurses–Canada–Biography

SGBIP: Red Cross
Women
Canada

Should commercially-available subject cataloging such as this from *Subject Guide to Books in Print* be accepted "with little or no modification"? Subject cataloging like this provides virtually no "control" at all, and virtually no possibility of readers' recognizing such headings within meaningful relationships. Note that the LC subject-strings would all show up intelligibly within larger browse screens, displaying other subdivision-aspects of the same topics in immediate proximity.

Result of "little or no modification" in Subject Cataloging: Unreviewed Cataloging from Bibliographic Utilities

What about the non-LC subject cataloging available from bibliographic utilities–the kind that Svenonius and McGarry found to be inaccurate or incomplete half the time? Again, the evidence is anecdotal; most reference librarians and catalogers just don't have the time to do statistical studies like Svenonius/McGarry.

Cataloger Jan Herd gave me an example she described as "not unusual in the books I receive." The title of the work was *The Credit Repair Rip-Off: How to Avoid the Scams and Do It Yourself.* The subject headings supplied by the copy cataloging were:

1. Debtor and creditor–United States
2. Debt relief–United States

Ms. Herd wrote to me:

> The first heading is a "law heading" and classes in KF1501 according to a law cataloger here in [this division]. [Note this cataloger's immediate recognition of the need for a proper tie to be established between LCSH and LCC.] He stated it should not be used on this book since it is not in scope as a law book. The second heading is also not appropriate for this book since Debt relief refers to macroeconomics . . . country level debt relief, renegotiation, etc.
>
> I received the book . . . . I had to change the headings to:
>
> 1. Consumer credit–United States
> 2. Credit ratings–United States
>
> The book was classed in HG3756 which corresponds to Consumer credit by country.
>
> This type of wrong thinking in assigning subject headings is not unusual in the books I receive. . . . When we multiply this kind of work on a daily basis we are polluting our database rapidly. We need a library EPA to impose "environmental impact charges" on libraries contributing to the pollution.

Usually I don't write down examples of bad copy cataloging unless there's a compelling reason; I have many other things to be doing with my time, and I generally just have to rely on what catalogers provide. Often, too, by the time I discover that I've overlooked some good sources due to their not showing up under the right headings, the reader who needs the books has vanished. I did write down an example, however, that was brought to my attention two months ago. A colleague of mine who is a rare book and manuscript cataloger in a private collection found, to her dismay, that her own scholarship was undercut by inadequate copy cataloging accepted by LC.

Result of "little or no modification" in Subject Cataloging: Undercutting Overviews Needed by Scholars

Dr. Melissa Conway's book, *The Diario of the Printing Press of San Jacopo di Ripoli, 1476-1484: Commentary and Transcription* (Firenze: L. S. Olschki, 1999), was published last year; and recently she was given an advance copy of a review of the book that will appear in 2001 in the journal *Book Collector*. Most of the review is irrelevant here, but on one point its writer faulted Dr. Conway's historical survey for not being updated by a particular book in the field that, the reviewer says, she should have read. Conway had been monitoring the appearance of books in the relevant field by regularly checking LC's catalog for works under the headings that had been applied to a standard work that she did make use of, Christian Bec's *Les Livres des Florins (1413-1608).* The subject headings assigned to this book are:

Books and reading–Italy–Florence–History
Libraries–Italy–Florence–History–1400-1600
Libraries–Italy–Florence–Catalogs

Florence (Italy)–Intellectual life

The book she is criticized for overlooking is Armando F. Verde's *Libra tra le Paret Domestiche*; this work itself is a kind of supplement to an earlier work by Verde, *Lo Studio Fiorentino, 1473-1503*. Evidently the non-LC cataloger who created the record for the *Libri* book didn't look at its contents carefully, but simply assigned to it the one subject heading given previously to the *Studio* record:

Universita di Firenze–History

In other words, according to Dr. Conway (who is herself a professional cataloger), the *Libri* book does indeed cover the subjects of **Books and reading** and **Libraries in Florence**, but the subject headings that ought to have indicated this were never assigned by the non-LC cataloger. And LC accepted the one inadequate subject heading "as is."

The ultimate point is that a serious scholar relied on a subject search of LC's catalog to do the "heavy lifting" it is supposed to do: not just to give her "something" on her topic, but rather to provide an *overview of the range* of significant, relevant resources available. And inadequate copycat subject cataloging, accepted with no modification, undercut that goal.

I do not mean to suggest that Dr. Conway's career is threatened as a result of inadequate subject cataloging; on the other hand, she is not in an academic position requiring "publish or perish" output, to begin with, or favorable reviews of it. An academic whose tenure is on the line in a similar situation, however, may have much stronger feelings about a library catalog that is supposed to, but doesn't, do the "heavy lifting" that a serious scholar expects of it.

The Need for Quality Subject Cataloging

And so I must beg to differ with Ms. Thomas's rather abrupt dismissal of the value of quality cataloging, which simply cannot be taken "with little or no modification" from the existing pools of ever-decreasing professional work.[28] Copy cataloging of subject headings and class numbers–if it is truly going to help library catalogs accomplish what scholars *need* to have accomplished–does indeed have to be checked with an eye to consistency, completeness, relationship, and accuracy. I realize, of course, that if Ms. Thomas is still promoting an opposite view in the wake of the Svenonius/McGarry study, and in the wake of the exposure of the factually false premises of the Mandel/Gregor article that she unquestioningly accepted as "knowledge," then nothing added here is likely to change her mind. But I sincerely hope that other participants in this Conference will realize that good subject cataloging–precoordinated, browse-displayed, linked to LCC, cross-referenced, and at specific levels–does indeed make all the difference in the world when its goal is understood to be that of providing *structured overviews of the range of significant sources relevant to a topic, rather* *than just "something"*–i.e., rather than just isolated and unintegrated information.

I'll say it again: If we as professionals are not making *knowledge* more available than it would be without our efforts–knowledge in its largest possible frameworks of relationships, interconnections, and linkages–rather than just isolated bits of *information*, then we are not fulfilling the most important responsibilities we have to our larger culture.

1. Mortimer Adler, *Ten Philosophical Mistakes* (New York: Macmillan. 1985), xiii.

2. LC itself has closed stacks, at least under its current administration; but most libraries using LCSH and LCC have open stacks in which this information would be immediately useful.

3. Note that Lois Mai Chan's Faceted Application of Subject Terminology (FAST), discussed in her "Expoiting LCSH" paper at <http://lcweb.loc.gov/catdir/bibcontrol/chan.html>, would, if applied to LCSH in both Web and OPAC environments, simply destroy the linkage of such strings to definite LCC stack areas. The same LCSH system, in other words, could not be used in both environments without great damage being done in the OPAC context, because postcoordination of the geographical "space" elements would destroy the indexing significance of the ordered string's link to LCC.

4. Unfortunately, the need for maintaining subject-classified bookstacks themselves seems to have dropped off the radar screens of many writers in our field. The continuing need for such classified shelving, and the reasons that it cannot be replaced by searching by class numbers within computer catalogs, are discussed at length in my paper, "Height Shelving Threat to the Nation's Libraries" at <http://studentorg.cua.edu/slislab/shelving.htm>. It also contains a discussion of the false notion that an "evolution" to digital forms is "inevitable." (In subsequent developments at LC, the matter seems to have gone into hibernation; the threat is no longer immediate.)

5. Numerous other examples can be found in the same book, as well as in the subsequent *Oxford Guide to Library Research* (Oxford U. Press, 1998).

6. Again, the FAST agenda (cf. note 3 above) would destroy such networks of cross-references if a scheme usable for LCSH in the Web environment were simultaneously forced onto LCSH in the OPAC environment. Since two separate LCSH systems cannot be reasonably maintained, the value of any proposed improvement needs to be critically examined for its impact in *both* environments. One hopes Ms. Chan's forthcoming study will address rather than ignore this crucial issue.

7. The evidence is not strong enough to establish a direct cause-and-effect relationship, but the observations made in a recent *Washington Post* article (April 26, 2000) by reporter Linton Weeks are not such that librarians and information professionals can simply ignore warning signs that are all around us, such as: "In the August 1999 issue of Conservation Biology, David W. Orr, a professor at Oberlin College, wrote that the human vocabulary is shrinking. By one reckoning, he observed, the working vocabulary of 14-year-olds in America has plummeted from 25,000 words in 1950 to 10,000 words

today. 'There has been a precipitous decline in language facility,' says Orr. 'This is nothing less than a cultural disaster.'" Weeks also quotes Keith Devlin, identified as dean of science at St. Mary's College in California and a senior researcher at Stanford; according to Devlin, "We may be moving toward a generation that is cognitively unable to acquire information efficiently by reading a paragraph. They can read words or sentences–such as bits of text you find on a graphical display on a Web page–but they are not equipped to assimilate structured information that requires a paragraph to get across. . . . Half a century after the dawn of the television age, and a decade into the Internet, it's perhaps not surprising that the medium for acquiring information [that a large number of the 10,000 college students surveyed] find most natural is visual nonverbal: pictures, videos, illustrations and diagrams." The dumbing down of learning–the loss of larger knowledge frameworks in our culture–is also commented on by Vladimir N. Garkov, "Cultural *Or* Scientific Literacy?," *Academic Questions*, 13, 3 (Summer, 2000), pp. 63-64: "A report on the first national assessment of our 17-year-old students' knowledge of history and literature found that this 'nationally represented sample of eleventh-grade students earns failing marks in both subjects.' A more recent study on cultural literacy, reported in the *Chronicle of Higher Education* (14 June 1996) found that only 7 percent of our graduating college students answered fifteen or more of the twenty questions correctly. The results from the National Assessment of Educational progress history exam show that only four out of ten high-school seniors demonstrated even a rudimentary knowledge of their own American history." Garkov cites Diane Ravitch and Chester E. Finn, Jr., "What Do Our 17-Year-Olds Know? A Report on the First National Assessment of History and Literature (New York: Harper & Row, 1987); Study on cultural literacy, *Chronicle of Higher Education*, 14 June, 1996; and L. Hancock and P. Wingert, "A Mixed Report Card," *Newsweek*, 13 November, 1995, 69.

8. Walt Crawford and Michael Gorman, *Future Libraries: Dreams, Madness, and Reality* (Chicago: American Library Association, 1995), p. 5; emphasis in original.

9. Martin Dillon, "Metadata for Web Resources: How Metadata Works on the Web." <http://lcweb.loc.gov/catdir/bibcontrol/dillon_paper.html>

10. F. W. Lancaster, "Second Thoughts on the Paperless Society," *Library Journal*, 124, 15 (September 15, 1999), 48-50.

11. Walt Crawford, "Paper Persists: Why Physical Library Collections Still Matter," *Online*, 22, 1 (1998), 42-48.

12. Dillon, "Metadata" (ibid.).

13. Lancaster, *ibid.*

14. State-of-the-art or overview "review" articles are especially prized by researchers. But it takes reference librarians to point out both the very existence of such articles, and the ways to find them.

15. Mr. Dillon's book *Interfaces for Information Retrieval and Online Systems* (New York: Greenwood Press, 1991) contains the following notice:

> "All rights reserved. No portion of this book may be reproduced, by any process or technique, without the express written consent of the publisher."

Lois Chan's books are similarly frozen in non-shifted formats; both her *Guide to Library of Congress Classification* (Englewood, CO: Libraries Unlimited, 1999) and her *Library of Congress Subject Headings* (Libraries Unlimited, 1995) contain identical boilerplate:

> "No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher and the author."

My own books have similar notices. Since the current copyright law protects such works for the life of the author plus seventy years, none of these books is likely to make "the shift" at all. And *should* some of them actually become digital, they will still not be accessible from anywhere, at anytime, by anyone on the Web; their digital versions will likely have physical-place use restrictions not appreciably different from their print counterparts.

16. The figure comes from RLG's Walt Crawford, in an email to me.

17. Lois Mai Chan, "Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources," <http://lcweb.loc.gov/catdir/bibcontrol/chan.html>

18. Lois Mai Chan and Theodora Hodges, "Entering the Millenium: A New Century for *LCSH*," *Cataloging & Classification Quarterly*, 29, 1-2 (2000), 225-34.

19. Regina Reyolds, "Partnerships to Mine Unexploited Sources of Metadata"; and Priscilla Kaplan, "International Metadata Initiatives: Lessons in Bibliographic Control," both available through <http://lcweb.loc.gov/catdir/bibcontrol>.

20. Caplan, Ibid., p. 6.

21. Findable at <http://lcweb.loc.gov/catdir/bibcontrol/thomas.html>

22. There are large commercial and governmental forces at work to get ordinary citizens connected to the Internet *in their homes*. Businesses promote home access because it enables them to target specific audiences and market groups, and to reach them (and their credit cards) immediately and interactively. Government, too, sees civic and educational goals being fostered by the same household hookups to the Net. In remarks made in December of 1999 in the Rose Garden, President Clinton noted the recent successes of public-private partnerships in closing the "digital divide" by wiring all schools and classrooms to the Internet. But he then went on to add, "there's still a lot more to do. We must connect *all of our citizens* to the Internet *not just in schools and libraries, but in homes*, small businesses, and community centers" [emphasis added]. Two months later, in announcing a multi-billion dollar federal program to solve the problem, he said, "Our big goal should be to make connection to the

Internet as common as connection to telephones" (*Washington Post*, 2/3/2000, p. B04). This is a politically popular agenda that will probably be pursued by whoever succeeds Mr. Clinton.

23. Dorothy Gregor and Carol Mandel, "Cataloging Must Change!," *Library Journal* (April 1, 1991), 42-47.

24. Sarah E. Thomas and Jennifer A. Younger, "Cooperative Cataloging: A Vision for the Future," *Cataloging & Classification Quarterly*, 17, 314 (1993), 237-57.

25. Thomas Mann, "'Cataloging Must Change!' and Indexer Consistency Studies: Misreading the Evidence at Our Peril," *Cataloging & Classification Quarterly*, 23, 3-4 (1997), 3-45.

26. Ibid., pp. 37-39.

27. Elaine Svenonius and Dorothy McGarry, "Objectivity in Evaluating Subject Heading Assignment," *Cataloging & Classification Quaterly*, 16, 2 (1993), 5-40.

28. Ann Huthwaite notes in her paper, "At the same time that this revolution has occurred there has been growing pressure on publicly funded institutions to reduce costs. Libraries throughout the world have been cutting back on expenditures and services." ("AACR2 and Its Place in the Digital World," <http://lcweb.loc.gov/catdir/bibcontrol/huthwaite.html>, p. 2.) Is there any doubt that more and more cataloging is being relegated to technicians?

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Matthew Beacom

Catalog Librarian for Network Information Resources
Sterling Memorial library
Yale University
130 Wall St.
New Haven, CT 06520

Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources

About the presenter: Matthew Beacom has been a cataloger and a librarian for 10 years. He catalogs networked information resources for Yale University Library. Prior to his current position, Beacom cataloged books for the Beinecke Rare Book and Manuscript Library at Yale. He has been a member of ALCTS CC:DA (1996-2000) and a member of the PCC Standing Committee on Automation (1999-2001) and was once chair of the LITA/ALCTS Interest Group on Technical Sevices Workstations (1998). Beacom is currently a member of the ALTS CCS executive committee (2000-2002). Among the specific tasks he has worked on for professional committees are the ALCTS CC:DA Task Force on Harmonization of ISBD(ER) and AACR2 [1998-99] and the PCC SCA Task Group on Journals in Aggregator Databases.

Full text of paper is available

Summary:

The advent of the World Wide Web has initiated profound changes in how

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

we use information. For librarians and those whom we serve, the most important changes may be in how new knowledge is created, how it is packaged, how it is published or disseminated, how its use is controlled, how it can be found, used, and saved for later use. In response to the Web and the cultural changes associated with it, librarians are thinking anew about how we enable people to have access to sources of information and knowledge. We are radically re-examining cataloging and catalogs.

In this paper, I address four problems or rather four changes in how we use knowledge that the library community must respond to. I address each change from the perspective of one who asks what the relation is between this change and AACR2, between this change and how libraries and librarians enable others to gain access to sources of information and knowledge. The four changes are:

1. The change in how knowledge is packaged,
2. The change in how knowledge is published and disseminated,
3. The change in how access to knowledge is controlled, and
4. The change in how we help others use knowledge as it is coming to be packaged, published, and restricted as networked resources.

Changes to how knowledge is packaged as a networked resource encounter AACR2 most immediately in rule 0.24. These changes profoundly affect how we understand and resolve the relation between content and carrier and greatly multiply the scale of the multiple versions problem.

Changes to how packages of knowledge are published and disseminated encounter AACR2 most obviously in the publication area, but their impact is not limited to this area. Networked resources are packaged in new ways. E-journal aggregations are one. After 400 years, will journals continue to be the dominant delivery mechanism for articles? Networked resources have new qualities. They are egregiously updateable. For networked resources, updateable publications may become the dominant pattern.

Changes to how access to packages of knowledge is controlled encounter AACR2 in the note area. With networked resources, access restrictions are commonplace. With networked resources, the mix of universal and local information in bibliographic records is shifted toward local information. URLs for licensed materials demonstrate the importance of this shift.

Changes to how we help others use packages of knowledge encounter AACR2 in its heart of hearts, in the role of the catalog record and the catalog itself as intermediaries between the resource and the user, the book and its reader. In a networked environment, the distance--space and time--between the catalog record and the resource is annihilated. A catalog on the Web is a portal to the Web.

IFLA's *Functional Requirements for Bibliographic Records* defines four user needs: to find, to identify, to select, and to obtain. To these four we must add a fifth for networked resources: to use. For networked resources, the catalog must deliver the resource to the user in ways that enable the user to make use of it to teach, do research, publish, etc. A catalog on the Web ideally delivers networked resources to the user's virtual workspace, a set of tools that enables the user to manipulate the resource--text, images, sounds, or data-- and put it to their own uses.

Glenn Patton,

commentator

OCLC, Inc.
6565 Frantz Rd.
Dublin, OH 43017-3395

About the commentator:

Glenn Patton is Manager of Cataloging Products Department in the Product Management and Implementation Division at OCLC. He has spent nearly 20 years doing support, training and product development activities for OCLC Cataloging services and products, including a major role in the redesign and implementation of the OCLC Cataloging Service. He and his staff are responsible for the implementation of online and offline cataloging-related products and services and for quality control activities related to WorldCat.

He serves as OCLC's liaison to the ALA ALCTS Committee on Cataloging: Description and Access, to Online Audiovisual Catalogers and to the OCLC CJK Users Group. He is a member of the Program for Cooperative Cataloging's Standing Committee on Training and the IFLA Standing Committee on Cataloging. He has also served as a member of the MARBI Committee.

Prior to coming to OCLC in 1980, he spent 11 years as Music and Fine Arts Librarian at Illinois Wesleyan University, Bloomington, Illinois. He holds B.Mus. and M.A. degrees from the University of Kansas and an M.S.L.S from Columbia University.

Full text of commentary is available

Library of Congress
May 9, 2000
Comments: lcweb@loc.gov

Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources

Matthew Beacom

A paper for the conference "Bibliographic Control for the New Millennium" held in Washington, DC at the Library of Congress, November 2000.

Final version

INTRODUCTION

The advent of the World Wide Web marks profound changes in how we use sources of information such as databases, indexes, and archives and how we use representations of knowledge such as maps, pictures, sounds, books, and journal articles. We are using the Web to change how we communicate with one another-how we read and write, how we speak and listen. We are using the Web to change how we do business-how we make and assemble things, how we buy and sell things. We are using the Web to change how we entertain ourselves. We are using the Web to change how we work. In short, we are using the Web to change our culture, to change how we live.

For librarians and those whom we serve, the most important changes in technology and society may be in how such sources of information or representations of knowledge are created and used. Specifically, these are changes in how knowledge is packaged or represented in an enduring physical form, how those physical forms or packages of knowledge are published and distributed, how their use is controlled or restricted, and how these packages of knowledge can be found, used, and saved for later use. These changes are profoundly affecting catalogers, catalogs, and catalog users.

A few of the chief topics of concern include "content versus carrier," "multiple versions," the purposes of the catalog, and its role in a networked information environment. By "content versus carrier" I refer to those problems or issues that affect how we relate the intellectual or artistic content of a representation of knowledge to the carrier or physical format that embodies it when there is a one to one relation between content and carrier. By "multiple versions" I refer to same issues of that relation between content and carrier when the relation between content and carrier is one to many.

In response to the Web and the cultural changes associated with it, librarians and their peers are thinking anew about how we enable persons to have access to sources of information and representations of knowledge. To that end, we are re-examining cataloging, catalogs, and our own role as intermediaries between those objects that embody information or knowledge-what Arlene Taylor calls "information packages" in her 1999 book, *The Organization of Information*-and those persons who would use them, our patrons, readers, or users (p. ??).

Librarians and their kin have been actively responding to the Web. Their responses include the following. MARBI changed MARC to accommodate access to remote electronic resources. IFLA published *Functional Requirements for Bibliographic RecordsAACR2* chapter 9 (on computer files) with the *ISBD(ER)*. The Joint Steering Committee (JSC) sponsored the International Conference on the Principles and Future of AACR in 1997 to explore how we might fundamentally revise the cataloging rules. Among the follow-up projects from that conference are a logical analysis of *AACR2* by Tom Delsey; a set of recommendations relating to issues of seriality in AACR by Jeans Hirons, et al.; and revisions to *AACR2* rule 0.24 that may establish the primacy of content over format for catalogers and user by CC:DA and the JSC. In light of all this thought and action, what problems of networked resources and *AACR2* could possibly be called unaddressed? The issues I discuss here, thus, are not literally unaddressed. For many others have considered, written, and acted on them. We are not, however, finished thinking about and acting on these issues.

In this paper, I discuss four changes in how we use sources of information or representations of knowledge, briefly evaluate the magnitude of these changes, detail some connections to *AACR2*, and suggest a few changes to *AACR2*. I address each of these four changes from the perspective of one who asks what is the relation between these changes and *AACR2*, between these changes and how libraries and librarians enable others to gain access to information and knowledge?

The changes

- From tangible to intangible media: how sources of information and representations of knowledge are manifested or packaged on the Internet
- From books and journals to services and databases: how such knowledge packages are published and distributed on the Internet
- From buying to leasing: how access to knowledge packages is controlled on the Internet
- From ascertaining to using, a new purpose for the catalog: how we help others to use knowledge packages on the Internet

Following the discussion of these four changes, I make 12 recommendations for changes to *AACR2.*

The size of the changes

Before discussing these changes individually, let's turn to estimating the magnitude of these changes. How big are these changes? Do they matter a little, a lot, or do they change everything? Christine Borgman, in her fine book, *From Gutenberg to the Global Information Infrastructure*, wisely distinguishes continuous or evolutionary patterns of change from discontinuous or revolutionary patterns of change. Her analysis of the technological and social change we are living through leads her to a reasonable view that she calls "co-evolutionary." (p. ??) Her view emphasizes the mixed and dynamic nature of the various responses of people and the organizations to technological change and the unanticipated consequences of such responses on further technological and social change. So how big are these changes? Overall, I argue they are radical changes that will in time transform how we create and use knowledge packages. To play on Borgman's vocabulary, let's call it "co-revolutionary." It is not an absolute break with the past, but it is changing everything. We are crossing a digital divide.

THE CHANGE FROM TANGIBLE TO INTANGIBLE MEDIA

On one side of this digital divide we have such traditional media as books and journals for texts, photographs and films for images, records and tapes for recorded sounds. All are *tangible* objects that contain or carry information or knowledge-the intellectual or artistic content. As librarians we are very familiar with this. Catalogers who work with knowledge packages that do not fall into simple types of materials know this junction of content and format as the content versus carrier problem. The things we catalog, what I'm calling knowledge packages-books, online databases, maps, e-journals, recorded music, digital archives, journals, corporate Web sites, etc.-are complex objects. They are mixtures of information or knowledge and physical format or carrier. Until a short time ago, everything we knew about knowledge packages was based on our experiences with traditional-*tangible*--media.

On the other side-the one we are crossing to-we have the new media, the Web, the Internet. The new media are *intangible*, untouchable. We still have such familiar kinds of content as texts, images, and sounds, but with a difference. Our experience of the texts, images, sounds, etc. carried by the new media are at one remove from ourselves-mediated by our computers. We can do wonderful things with the new media, but at a price. We can no longer touch them. We based *AACR2* on the idea of cataloging the item in hand. And now the things-knowledge packages-we catalog can't be handled.

Yet things on the Web are still things, after all. The things we catalog need not be tangible to be *things*. The documentalists, who flourished earlier in this century, understood this. Michael Buckland in his 1991 article, "Information as Thing" quotes, Briet's definition of a "document" as any concrete or symbolic indication of a physical or mental phenomenon that has been recorded for reconstructing or recreating that phenomenon." (p. 355) Although similar in vocabulary to the Internet era idea of document-like objects, it is actually the reverse. Contrary to what their name implies, documentalists were expanding the idea of document to include all representations of knowledge within a comprehensive concept. The idea of document-like objects as applied in the Dublin Core initiative restricts its scope to those representations of knowledge that have the qualities or characteristics of documents.

Knowledge packages on the Web do not lack physical qualities. They lack some familiar qualities and have some we are not used to. Web things--changeable, ephemeral, and adaptable--can be lost, found, or hidden, bought, sold, or leased, corrupted, destroyed, or preserved, known, cited, or used. These Internet-based knowledge packages retain many recognizable characteristics. For example, they are still intellectually recognizable as particular types of knowledge packages such as reports or poems, drawings or music, census data sets or road maps. The move from touchable to untouchable media fundamentally alters how we act on and use knowledge packages. *AACR2* defines some rules on one particular use of knowledge packages-the creation of metadata surrogates or catalog records for use in library catalogs.

The change to how representations of knowledge are packaged-the change from tangible to intangible media--encounters *AACR2* most immediately in rule 0.24, but also in the physical description area (ISBD area 5), and in the materials specific details area of *AACR2*. How networked resources are packaged profoundly affects how we understand and resolve the relation between content and carrier. The potential for multiple packages of the same or nearly the same content in networked resources greatly increases the scale of multiple version problems. These problems are central to bibliographic control of networked resources.

A cardinal principle

Let's look at *AACR2* rule 0.24 in more detail. Before this digital divide, we based description on the "physical" form of the item in hand. But by physical form we meant the categories referred to by chapters 2 to 12 of *AACR2*, the types of material. Despite the mixing of intellectual and physical formats, this arrangement worked well for the most part. Important exceptions include serials and reproduction microforms. The issues or problems with serials and reproduction microforms are well known. The treatment of seriality in *AACR2* is being significantly but not *radically* revised now. Reproduction microforms and the larger issues of reproductions and multiple versions in the digital environment have yet to be successfully addressed. The discussions now underway in the cataloging community including a draft LCRI 1.11A Non-Microfilm and Electronic Reproduction is unsuccessful primarily because of our failure to define what an electronic reproduction is. Because of the ease with which networked resources can be re-purposed, multiple version issues regarding bibliographic control and user services are approaching crises.

Content and carrier

Rule 0.24 has been revised now so that catalogers are directed to bring out all the characteristics of the item being described (i.e. formats detailed in *AACR2* chapters 2-12.) This is a solid improvement. It represents a big response to the change from tangible to intangible resources, but it is not the radical change that we will soon need.

There is an element of abstraction to all things. The saying "the eyes see what the mind knows" is a testament to the mixing of physical and mental in what we call things. A book is an *idea* as much as it is a tangible object. A Web site or e-book is also an idea as much as it is an intangible object. The abstraction level is just a little higher for us because we can't touch an e-book or a Web-site. Current practice is mixed for both analog and digital media. 0.24 classically says to base the description on the item in hand, in effect catalog the manifestation. But our practice is mixed. For example, the Library of

Congress's microform practice vs. *AACR2*-conformant practices such as that at the National Library of Canada. Or, the CONSER single record practice-a form of dashed-on entry for the 21st century-and the separate record practice.

Content and carrier cannot be separated without breaking the link between the work and the item. That link is vital to successfully serving the user who needs to have some particular item that is the right work, the right expression, and the right manifestation. If we can't break the link between the work and the item without harming the user, what do we do?

The distinction that we need to make is between analog and digital, stand alone and networked, tangible and intangible. By using the types of material defined in chapters 2-12 of *AACR2* as primary types we continue to confuse types of carriers with types of content and with modes of publication over time. Delsey's suggestion that *AACR2* be reorganized by ISBD area is a powerful idea. The test case reorganization done by Library of Congress's Cataloging Distribution Service unit shows that this reorganization is not trivial or mechanical. Serious intellectual effort to reorganize *AACR2* by ISBD (or ISBD-like) area is needed now. This approach may also be described as creating a super chapter 1 of *AACR2*. Practicing catalogers need guidelines for cataloging traditional types of materials or other categories of knowledge packages. These guidelines must be based on the rules, but need not be part of the rules themselves. In other words, instructions for cataloging some of the materials now in *AACR2* chapters 2 to 12 may best be separately published as material specific guidelines and not rules.

Multiple versions

How do we address multiple versions or multiple formats at the level of the records or surrogates that we create for our catalogs? For the multiple version issues, the big question is what things should our records be surrogates for? Should we make surrogates or records for the content itself abstracted from its carrier or should we make surrogates for the content *and* its carrier? In the terms of the *Functional Requirements for Bibliographic Records*, the question is should we catalog each expression of a work or each manifestation of the work? In current practice and rules, we mix content and carrier in flat surrogates or records. In the terms used in Functional Requirements for Bibliographic Records, each catalog record we make for a given knowledge package generally mixes information about the package at four levels: the work, the expression, the manifestation, and the item. Imagine a text in 6 formats: XML, SGML, pdf, postscript, on a DVD, and in print. Should that be described in 1 surrogate, 2, 3, or 6? If 1, how are the manifestations articulated. If more than 1, how is the principle of division made clear? If 6, how are the bibliographic relations among the 6 manifestations described? Further, by what criteria will we decide? If the content is the same, shouldn't it all be on one record? That is, in effect, the notion of cataloging the expression rather than each manifestation. But is content the only relevant criterion?

The chief contra-argument to cataloging the expression is that content is *inescapably* joined to carrier. This union of content and carrier makes the knowledge packages that our users actually use. The combination of content and carrier is what makes knowledge *usable*. These knowledge packages are the objects around which libraries-servers, buildings, networks, staffs, services, collections, and purposes-are built. Libraries collect things, knowledge packages that mix item, manifestation, expressions, and work aspects into a whole. To meet its purposes, the library catalog needs to make those things the library has accessible to users. It does not matter whether they are on a shelf or on a server. The focus of on the thing itself is still vital. *Representations of knowledge are things.* When we lose sight of this fact, we lose our way.

Records for knowledge packages in multiple formats

Some catalogers want to split up knowledge things-describing each manifestation separately. Others want to lump knowledge things together-describing all manifestations of an expression together. Each group thinks its way is best, especially best for users. But the dichotomy of splitters and lumpers is a false one. It is not a matter of either/or. It is a matter of when and where. The criterion is what can we do that serves the user best.

The classic example of this split *and* lump approach is described in the *Guidelines for Bibliographic Description of Reproductions*. A tiered record approach both splits and lumps. One tier describes the expression (and provides work level access points), another tier describes the manifestations (and provides manifestation level access points.) This powerful model deserves renewed interest and effort. This model would solve our multiple version problems. However, it is not without problems. Among the best known are the need for compatibility with older records, with MARC formats, and with legacy OPAC and bibliographic utility systems. The need for backward compatibility may be one of the strongest pulls on librarians to adapt *AACR2* rather to jump to a born digital metadata scheme like the Dublin Core. We have an installed user base that we don't want to and can't abandon, and we have institutional commitments to servicing analog materials.

There are other ways to split and lump. We could split at the point of record creation and lump at the point of display. Our rules could dictate that we split at the record level by cataloging each manifestation, and lump at the display level by linking each manifestation record into a integrated display of expression level and manifestation level information. Of course, this approach has its own disadvantages. Our OPACs would have to intelligibly and flexibly show bibliographic relationships among records and link (conceptually and mechanically) across records. Our records would need subtle and robust areas for managing relationships and linkages. The Web just happens to be a really suitable environment for doing both of these things. It may be far more possible to do this in the next 10 years than it was in the last 10.

CC:DA's recent recommendation on 0.24 to the Joint Steering Committee (JSC) takes another approach. Split sometimes and lump other times; make the choice based on a list of major/minor changes that are to be appended to *AACR2*. The most recent CC:DA recommendation to the JSC builds on the earlier recommendation by devising a list of major/minor changes that would guide catalogers in deciding when to create new records. This approach will not work. The list of changes is not the tool catalogers need. The tool we need is a coherently conceived record structure, such as the tiered or linked record structures mentioned above. The question cataloger's need to ask is not-- when do I make a new record? The question to ask is how do we effectively distinguish and display work, expression, manifestation, and item level information to users.

THE CHANGE FROM BOOKS AND JOURNALS TO SERVICES AND DATABASES

Changes to how packages of knowledge are published and distributed encounter *AACR2* most obviously in the publication, distribution, etc. area (ISBD area 4). But the impact is not limited to this area. Indeed, the impact of these changes is far broader than conventions for recording the places, names, and dates associated with publishing.

I address six aspects of this change and its affect on cataloging.

- What does "published" mean on the Internet?
- What are the consequences of the Internet flood of information sources?
- How does reference linking change cataloging?
- What new means of publishing and distributing representations of knowledge are likely to dominate the Web?
- What is the future for books in a media environment so conducive to interactive multimedia and continuous updating?
- What is value to users of imprint information (ISBD area 4) in the digital era?

What does "published" mean on the Internet?

Changes in what it means to publish affect *AACR2* fundamentally. The change from traditional media to digital networked media disrupts our understanding about what is and is not published, about what it means to publish. The dictionary says publish means "to make generally known," "to place before the public." This is fundamental, but only part of what it means to publish.

Our understanding of publishing is complex. Publishing is an intellectual, social, economic, and technological phenomenon. Our understanding is tied to the central distinction between public and private spheres of life and blurred by phenomena such as gray literature and invisible colleges. The Internet and our particular uses of it are driving changes in the public and private distinction, greatly increasing the visibility of gray literature, and through peer to peer networking making invisible colleges on a global scale possible.

Our understanding of publishing is more fully developed through the concepts expressed in words like original, copy, edition, impression, and reproduction. These concepts are traditionally associated and frequently used in our work. These familiar and traditional concepts have been built on our experience of analog formats. As we cross the digital divide, we extend these concepts to digital, networked knowledge packages or things in order to keep control of the new materials. We make do, innovate, and adapt.

In some library and publishing ventures these adapted working definitions have been useful. The re-publications of journals and books by JSTOR and netLibrary are examples of publishers extending the use of these familiar publishing concepts to digital networked materials. Other ventures such as pre-print databases, e-journal aggregations, and personalized services that may replace traditional textbooks such as those offered by MetaText suggest new modes of packaging and distributing recorded information and knowledge.

Will these extensions or adaptations be for naught? In the context of the Internet, applying such familiar concepts of "original," "copy," "edition," "impression," and "reproduction" is often of doubtful value. For example, in the analog age the number of copies of a knowledge package made for distribution is limited. In the digital age, copy is more likely to be a verb than a noun. This is a small but telling difference. In the analog age, these terms represent fairly precise concepts; in the digital age, they become metaphors, new parts of speech, or anachronisms.

Since the Internet makes it so easy "to place things before the public," some have argued that we should treat everything on the Internet as published. However, the published or not published division can be made and may need to be. *Digital Dilemma: Intellectual Property in the Information Age* raises these issues clearly and makes a strong case that on this side of the digital divide publishing will still be a complex and nuanced phenomenon. We may include on the not published side such things as author's drafts, notes and other materials not used in finished products, as well as "private" material like e-mail or calendars or digital diaries. In revising *AACR2*, we need to decide how we will make and use distinctions between published and unpublished digital networked materials we may add to our collections and catalog.

For example, we may wish we could treat all online manifestations of some content as reproductions of an analog original form and use LC's microform practice to guide our cataloging. On the other side of the digital divide, print is dethroned. Print becomes just another output option, one that can be invoked or not invoked by the publisher, a wholesale reseller, a retailer, a library, or a reader. If there is an "original," it is online. We need to develop a new vocabulary and new concepts out of our analog *and* our digital experiences. And we need to use these ideas in *AACR2*. Work, expression, manifestation, and item have already been mentioned. These terms and ideas take us a long way. Problems with our concept of reproductions have also been mentioned and it requires further work.

The Internet flood of information sources?

A more pressing consequence of the increased ease of publishing is the sheer volume of materials on the Web. We are experiencing, in part, a tidal wave of gray literature. (Another portion of the wave is the result of a global village effect, e.g. every town's newspaper is online and available at any computer.) On the Internet, the distinction between published materials and gray literature is weakened. So much gray matter is so easy to find on the Web that more formally published material is lost or obscured. (Much of that formally published material is also hidden behind access restricting checks.) Metadata developments like the Dublin Core are partially predicated on this blurred distinction. Too many people are making too much material public through too many channels or outlets for traditional methods of bibliographic control such as library catalogs or national bibliographies to suffice. In revising *AACR2*, we need to decide what relation library catalogs will have to the Internet. I address this aspect more fully below in the section on the changing role of the catalog, but the key word is *selection*.

The catalog and reference linking

Content can be re-packaged and leased many times to many such groups because the Internet makes it easy for many different agencies to license the same content to many different groups. This is one source of the proliferation of knowledge things and the records that describe them. Clearly, this can be a collection development issue for libraries-how may times do you want to buy the same content for your user group? But for the makers and users of catalogs, a defining aspect of networked resources adds a twist to the multiple version issues. URLs are not universal. The URL that links the resource described in one record only works for members of the licensed user group. This is no surprise to many and is one reason why MARBI defined the electronic location field (MARC tag 856) in both the bibliographic and the holdings formats.

MARBI has addressed the need for records or surrogates of online resources to link directly to the resources themselves. *AACR2* has not. URLs are often consider something like a shelf location or other completely local information and thus outside of any concern by *AACR2*. Without creating rules for making hyperlinks from surrogates to the resources themselves, *AACR2* breaks a linkage that is a defining characteristic of networked information. Where should this kind of information be: in notes, in standard numbers, a new section, or in a general rule and then added throughout the code as needed? I recommend the later. *AACR2* must explicitly address hyperlinks such as URLs, URNs, and others in the cataloging rules. ISBD needs to address this too. Otherwise in a digital world, the makers and users of digital media will ignore *AACR2* (and the ISBDs). Dublin Core is designed for Web-based knowledge packages. MARC has adjusted to the Web with the 856 field. The makers and user of *AACR2* (and the ISBDs) must recognize the critical importance of reference linking in a networked information environment. Redesign the rules to fit a publishing environment of pervasively inter-linked knowledge packages and a metadata environment of similarly inter-linked surrogates or records.

New publishing and distributing methods

Networked resources may be packaged in new ways. New bibliographic entities and new bibliographic relationships are native to the digital, networked environment. E-journal aggregations are one example. Are they convenient bundles of journals or are they precursors to new delivery mechanisms for articles? Article databases are replacing journals as the dominant *delivery mechanism* for articles. Journals are not likely to just go away. Their roles will change. They will continue to have powerful editorial functions with resulting value as brands and as a useful search limit term. Their function as devices for article delivery to the user will lessen in importance. Such a transformation in publication practices would significantly affect *AACR2* chapter 12 both in its current form (i.e. Serials) and in its emerging form (i.e. Continuing Resources.)

In general, the impact of new kinds of knowledge things on the Internet on *AACR2* is to undermine

AACR2's extensible structure. Although *AACR2* is designed to adapt to new formats of materials by adding chapters, adding new chapters for proliferating e-formats (tangible and intangible) is not a viable choice. Second, we have mixed up physical formats with characteristics and qualities. For example, seriality is a condition not a format. It is potentially applicable to any knowledge thing we can imagine: texts, images, cartographic information, sound recordings, etc. The current recommendations before the Joint Steering Committee now recognize this, but for many reasons, mostly practical ones, the changes are mainly contained within the chapter for the serials format, chapter 12. We are still trying to compartmentalize seriality, to treat it as a format not as a range of conditions or characteristics that might apply to any knowledge package.

Furthermore, digital networked resources are at least as likely to be *blends* of what we have traditionally called formats as they are likely to be, shall we say, single malts. Now we use adjuncts to *AACR2* like *Guidelines for bibliographic description of interactive multimedia and Cataloging Internet Resources: a Manual and Practical Guide* to retrofit the rules. We have made modest changes to the rules as in the revision of *ISBD(CF)* to *ISBD(ER)* and the recent efforts to harmonize *AACR2* chapter 9 with *ISBD(ER)*. Now, on this side of the digital divide, we need to rethink our rules with the networked environment as the technological and social base for communication. To accomplish this a thoroughgoing revision of *AACR2*, such as that suggested by Delsey is needed. We are making progress. We are moving quickly for our profession, but slowly for the larger networked environment in which we now find ourselves.

One specific impact may be seen in possible responses to the development of article databases. Two options come to mind. Return to article level cataloging. This is possible but unlikely to be a successful strategy. Our experience in the past century with third party journal article indexing has demonstrated its efficacy relative to cataloging journals article by article. A second option is to link cataloging and indexing information in ways that the user sees as seamless. Developments with reference linking tools like jake and SFX indicates the power of such a smart, scalable approach.

For *AACR2* to support such deep or integrated linking, the cataloging community needs to add a new area on linkage and relationships to *AACR2*. In *AACR2* hypertext or hyperlinks-and the technological and social environment that supports it and expects it-does not exist (except perhaps in chapter 9 and only grudgingly and implicitly.) It should be a fundamental principle of cataloging in a digital age that all records and other metadata surrogates should be designed to link to other surrogates that describe the same resource at different levels of granularity or other related resources (even those that use other metadata schemas. Such linkages can be applied to analog and digital materials. .

New resources, new qualities

Networked resources have new qualities. For example, digital networked resources are egregiously updateable. In *AACR2*, updateable publications such as looseleafs are marginal at best. For networked resources, updating publications may become the dominant pattern. Such resources are also open to combinations of multiple media (text, images, sound, etc.) in one publication in ways that are unthinkable in print or other analog formats. Furthermore, such resources are inherently linkable. One example of this shift may be a move from monographs to interactive services. This change may be indicated by NetLibrary's development of its MetaText product. This product is a set of Web-based communication and analysis tools with interactive multimedia content-enriched textbooks as its content. (NetLibrary)

The development and dominance of services over distinct objects may lead us beyond what a catalog can contain. Or perhaps it is only an issue of granularity. We can catalog the services as entities and not the shape-shifting products one can produce on demand from such services. This is similar to collection level cataloging. But how do we contribute to making materials below the level of the whole service accessible? This is a new responsibility for the catalog and for *AACR2*. The key is cross-profession collaboration and inter-linked metadata standards. We are doing that now with archival finding aids that use the EAD DTD. The archival collection can be cataloged using *AACR2* and expressed in MARC for transport and use in OPACs. The record includes a hyperlink to the EAD- encoded finding aid for that collection. Librarians use one standard for the catalog record and archivists use another for the finding aid. Users benefit from metadata created by two different but related support communities. The assumption of catalogers must be that the surrogates or records they create will be used in conjunction with other forms of metadata. *AACR2* revisions must explicitly declare this assumption, and it must design rules around its consequences.

The value of imprint information

In a digital networked environment what is the value to users of the information recorded in the publishing, distribution, etc. area? Does the Internet affect the value of the imprint information and its use? For example, does *place* of publication matter online? The place of publication may not matter at all, or it may matter in new ways. Users may find new values or new uses for imprint information. If the former, why record it? If the latter, will those new uses change how and what we will record? Imprint information may become more important for access than for description and identification. Citation practices may change from the conventions developed in an analog age. (National bibliographies may also change their practices as the Internet enables increasing globalization of enterprises like commercial publishing.) *AACR2* needs to address this particular question but also more generally ask what is the role of *transcription* in an era of networked resources. Transcription has never been and should not be an end in itself. It has always served the function of identification by enabling the surrogate to mirror the resource it describes.

This shift in the value of imprint information isn't just a matter of digital form. It is also a consequence of globalization. The named publisher may be little more than a brand name in a multinational media conglomerate, and places of publication have been proliferating for print publications, too. But there is an analog vs. digital divide here, too. In the analog era, it is used in part as evidence to identify the manifestation and the expression-the edition. In digital era, are such indirect indications of edition as useful? The *name* the publisher may matter to the user and the library less than the name of the e-seller, the e-aggregator or e-jobber? Electronic materials may be re-packaged by so many vendors that the publisher may not matter to users or the library as much as the vendor may. The agency with which the library or the user has signed a licensing agreement may be far more important for description and access in the digital era than the publisher.

THE CHANGE FROM BUYING TO LEASING

Changes to how one controls access to packages of knowledge currently encounters *AACR2* primarily in the note area (ISBD area 7). But its impact is not limited to this area. The change from an environment dominated by buying and selling knowledge things and controlled by copyright law to one dominated by leasing and controlled by licensing agreements is fundamental. It alters the relations between the library and the things it collects, between the library and its users, and between the knowledge things themselves and their use and usability. In the analog era, communication and scholarship ranged across a fair use commons. On this side of the digital divide, the fair use commons is being claimed and fenced in. The impact of these changes on *AACR2* is critical to the relevance of the rules within an environment where licensing agreements control the exchange of information.

Access restrictions are uncommon with many analog materials. A notable exception is archives of unpublished materials. In the analog era, copyright is the defining rights management paradigm for the relations between publishers and users. Habits of sale and use have developed in this relatively stable technological, social, and legal context. Copyright itself is an extrinsic context within with items are bought and sold. With networked resources, though, access restrictions are not only commonplace, but also vital characteristics of digital objects. Although they are not strictly speaking intrinsic qualities, access restrictions or, more generally speaking, rights management conditions are profoundly in-twinned with the *use* of digital networked resources.

Notes about access restrictions are helpful to users. Users might not read them, but they are better than nothing is. A combination of universal note and local note is often most useful. Without them the user has no hope of knowing what items in any search result are or are not accessible to them until *after* they attempt to retrieve each item. An access forbidden message will let them know at some point, say when they try to see the full text of a particular article. However, notes alone are inadequate to rights management in an environment dominated by leasing.

AACR2 has no area for dealing with rights management. The notes area and terms of availability section are inadequate substitutes for a rights management area. Access restrictions and rights management must be explicitly addressed by the cataloging rules. Users are ill served by surrogates that are not rights aware. The Dublin Core element set has led the way for the library community by making rights management one of its 15 elements. Publishers are developing their own metadata standard, ONIX International Release 1.1., that also includes rights management elements. (Editeur) A new area is required in *AACR2* to address the, for all practical purposes, intrinsic needs of digital networked knowledge things for rights management information specific to them.

THE CHANGE FROM ASCERTAINING TO USING: A NEW PURPOSE FOR THE CATALOG

Changes to how we help others use packages of knowledge encounter *AACR2* in its heart of hearts: in the role of the catalog record and the catalog itself as intermediaries between the book and its reader, between the resource and its user. The role of the catalog (and the records that populate it) is changing in two big ways. The first is the change from a finding aid to media delivery device to a virtual workspace. The second is the change from the premier research tool *in* the library to a valuable research tool in an Internet toolbox. The consequences of these two changes are far-reaching-they do change everything.

In a networked environment, the distances between the catalog record and the resource itself are annihilated. On the Web, the catalog record and the resource are hyperlinked together. They are not made one, but they are no longer independent and separated objects. The data and the metadata are physically *and* intellectually linked. A reader can use the surrogate to summon the resource itself. This sort of linkage has always been possible intellectually and imaginatively-quotations and citations are familiar examples of this. The Internet and hyper-linking make such intellectual or imaginative links real. Real in the sense that the links are physically present not just references and that the links can be used to make things happen-can bring the resource and the user together in a virtual workspace.

For networked resources, the catalog is not only a finding aid, a listing device. It is also multimedia delivery system. And it is more than that. A catalog on the Web is a portal to the Web. Like all such portals it is a door to a sub-set of the resources that populate the Internet. One library's portal may lead to a smaller or larger sub-set of resources than another, just as one Web search engine may index a smaller or larger sub-set than another engine. This is where the catalog is now, but it is not where it will stop. The development of the catalog will continue until it fulfils the promise of the fifth user need-use. The catalog must become a research tool that is integrated with the user's virtual workspace. The surrogates that populate our catalogs are no longer static and separate things. On this side of the digital divide, they are as dynamic and as linked-up as the resources they describe. This changes everything.

The Paris principles define the catalog as a tool for ascertaining whether or not some thing exists in a particular collection or collections. AACR1 is explicitly based on these principles. In *AACR2*, this basis is implied. Reduced to a single word, the purpose of the catalog is to *ascertain*. The *IFLA Functional Requirements for Bibliographic Records* defines the catalog in terms of meeting four user needs: to find, identify, select, and obtain. On this side of the digital divide, we must add to these four needs, a fifth: use.

For networked resources, display is insufficient. View, print, and save are only starting points. The catalog must deliver the networked resource to the user. Furthermore, it must do so in ways that enable the user to make use of the resource to meet the user primary needs. In an academic setting those primary needs are to teach, research, and publish. The catalog on the Web delivers networked resources to the user's virtual workspace, to the set of tools that enables the user to manipulate the resource-text, images, sounds, data, etc.-and put that content to their own uses.

Examples of such virtual workspaces that integrate data and metadata are now in use. NESSTAR, Networked Social Science Tools and Resources is one project developing such workspaces within the field of social science data archives. The NESSTAR project has developed sets of tools that allow researchers to identify, locate, download, and use data from sites on the Web. The system is built upon DDI (Data Documentation Initiative)-compliant metadata. (NESSTAR Web Site) Another, from the field of digital art images, is Luna Imaging's Insight software at the Visual Resources Center at Yale University Library. Yale Library's implementation of Insight is a collaborative experiment in digital access to resources from the Yale University library and museum collections to support classroom teaching in the field of Art History. (Browser Insight) The Insight tool is builds upon VRA (Visual Resources Association) Core metadata. Tools such as these are the future of the catalog.

The new purposes of the catalog require a new conception of surrogates and catalogs, one that supports the linkage between the surrogate record and the resource. *AACR2* can no longer ignore the new bibliographic world that hyperlinks and networks are creating. Letting the MARC format carry the load, leaves the rules for cataloging less than Web aware. Adding a new area to *AACR2* (and to the ISBDs) for linking information is not the best approach. Since hyper-linking within a networked environment is a pervasive aspect of communication and publishing on this side of the digital divide, trying to keep linking in one area is counterproductive. The MARC format is already expanding the use of URLs in fields beyond the 856. In practice catalogers are far beyond even those extensions. The new purposes must be explicitly addressed in *AACR2* and linking must be supported throughout the rules.

In the analog age, the catalog has been the premier research tool in the library. In the digital networked age, it is a valuable research tool in an Internet toolbox. *AACR2* has grown up and flourished in the relatively homogenous confines of the library and its collections, purposes, services, traditions, and community of users. Other institutions did similar things but often did them differently, for different purposes, for different people, and in different places. Art museums and galleries, archives and natural history collections, indexing and abstracting services, research labs and projects are among the more obvious peers of libraries in collecting, organizing, and keeping sources of information to serve their users. In the digital networked era, libraries and their peer institutions are no longer so isolated from one another. The Internet has created opportunities for collaboration and even competition where few had existed before. The Web offers a heterogeneous world of resources to the researcher. Library catalogs are only one tool in this wider world.

The defining role of a library is that it is a collection or collections of selected materials. This is true in the analog era and in the digital era. Everything else we know about libraries and what they do relates back to this fundamental act (and fact) of selection. What is different in the digital era is that this role must be made explicitly clear to the user and not implied by the traditional limits of tangible things, books, buildings, campuses, etc. The shift from implicit landmarks to explicit signs is a generally applicable effect of the move to a digital networked environment. One implication for libraries is that catalogs cannot serve users well if they are conceived of as stand alone systems, as portals to one library's selection of Web resources. The catalog must be integrated with other resource discovery tools. For example, users of a catalog must also be able to turn their search into a broader Web search.

One aspect of the opening up has been the phenomenal interest in metadata. Another has been the development of myriad crosswalks to enable one metadata format to be translated into another. A third should be changes to *AACR2* that reflect this new world order. In homely terms, *AACR2* has been an only child reared at home, and now *AACR2* has gone to nursery school and must learn to play well with others. Crosswalks are one way to play well with others, but they are exterior to the rules. Another way is to conceive of our rules as one way among many. We need to alter our rules and our principles so that we have the means to create records or surrogates that thrive within a rich, pluralistic world of dynamic and inter-linked resources and surrogates.

RECOMMENDATIONS

12 changes to make to *AACR2* to adapt it to a digital networked communications environment.

1. Change the purposes of the catalog by adding to the concepts in the Paris Principles those concepts of user needs expressed in the IFLA *FRBR*-find, identify, select, and obtain-and the fifth user need: *use*.
2. Change the concept of the catalog as a stand alone finding aid or listing device; explicitly state its ideal role as an tool designed to work well with other tools that use other metadata rules for their surrogates.
3. Change the orientation of display instructions from card production to online (hyper-linked) displays; change from editorial instructions to design guidance, include guidance for labeled and unlabeled displays, include explicit support for URLs and other reference linking techniques.
4. Use the concepts "work," "expression," "manifestation," and "item" as articulated in the IFLA report *Functional Requirements for Bibliographic Records* as a general framework within the rules. Concepts such as "edition," "impression," "original," or "copy" may continue to be highly useful for analog materials, but cannot be basic concepts of bibliographic control in an age of digital networked resources. Re-define the concept of a "reproduction" in an age of digital networked materials.
5. Thoroughly examine changing the arrangement of Part 1 of *AACR2* to follow an ISBD-like area order
6. Move instructions for cataloging particular types of materials out of the rules; collaborate with user communities to develop cataloging manuals (like *Bibliographic Description of Rare Books* and the *CONSER Cataloging Manual*) that are based on the rules.
7. Add new ISBD-like area for rights management information.
8. Add a new ISBD-like area for bibliographic relationships and reference linking.
9. Following the adoption of the proposed changes to chapter 12 (Serials), develop an ISBD-like area for the mode of issuance to include finite, serial, and integrating patterns of publication.
10. Eliminate *AACR2* chapter 9 (Electronic Resources); develop an ISBD-like area for the carrier aspects of all knowledge packages.
11. Further revise rule 0.24 so that the manifestation in hand or on screen remains the primary artifact being described; require that relations among manifestations or from manifestation to expression be articulated within or across records as needed.
12. Reconsider the role of transcription in descriptive cataloging. Since transcription is not an inherently suitable technique for describing dynamic or potentially dynamic resources, it may not be supportable as a primary means of creating identifiable surrogates.

REFERENCES

Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access (1995). Guidelines for bibliographic description of reproductions. Chicago: American Library Association

Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access (1994). Guidelines for bibliographic description of interactice multimedia. Chicago: American Library Association

Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access. Task Force on an Appendix of Major and Minor Changes (2000). Report [Online] Available <http://www.ala.org/alcts/organization/ccs/ccda/tf-appx1.pdf> [Oct. 11, 2000]

Association for Library Collections & Technical Services. Machine-Readable Bibliographic Information Committee (MARBI). Proposal 93-4

Buckland, Michael (1991). "Information as thing." Journal of the American Society for Information Science. 42(5):351-360.

Borgman, Christine (2000). From Gutenberg to the global information infrastructure. Cambridge, Mass.: MIT Press, 2000.

Caplan, Priscilla and William Y. Arms (1999)."Reference linking for journal articles." D-lib magazine. 5(7/8) [Online] Available http://www.dlib.org/dlib/july99/caplan/07caplan.html [Oct. 11, 2000]

Committee on Intellectual Property Rights and the Emerging Information Infrastructure (2000). Digital Dilemma: Intellectual Property in the Information Age. Washington, D.C.: National Academy Press [Online] Available http://www.nap.edu/books/0309064996/html/ [Oct. 11, 2000]

Delsey, Tom (1998). The logical structure of the Anglo-American Cataloging Rules-Part I. [Online] Available <http://www.nlc-bnc.ca/jsc/> [Oct. 11, 2000]

Delsey, Tom (1999). The logical structure of the Anglo-American Cataloging Rules-Part II. [Online] Available <http://www.nlc-bnc.ca/jsc/> [Oct. 11, 2000]

Dublin Core Metadata Initiative (1998). Dublin Core metadata element set, version 1.1: reference description. [Online] Available <http://purl.org/dc/documents/rec-dces-199809.htm> [Oct. 11, 2000]

EDItEUR (2000) ONIX International Release 1.1 [Online] Available <http://www.editeur.org/onix.html> [Oct. 11, 2000]

Gladney, Henry M. (1999). "Digital dilemma: intellectual property." D-lib magazine. 5 (12) [Online] Available http://www.dlib.org/dlib/december99/gladney.html [Oct. 11, 2000]

Hirons, Jean and Crystal Graham (1998). 'Issues related to seriality', in The principles and future of AACR : proceedings of the International Conference on the Principles and Future Development of AACR, Toronto, Ontario, Canada, October 23-25, 1997. Ottawa : Chicago :Canadian Library Association; American Library Association.

Hirons, Jean (2000). Revising AACR to accommodate seriality: rule revision proposals. [Online] Available <http://www.nlc-bnc.ca/jsc/ch12.htm> [Oct. 11, 2000]

Hirons, Jean (1999). Revising *AACR2* to accommodate seriality : Report to the Joint Steering Committee for Revision of AACR, with the assistance of ... the CONSER AACR Review Task Force, April 1999. [Online] Available <http://www.nlc-bnc.ca/jsc/ser-rep0.html> [Oct. 11, 2000]

International Conference on Cataloguing Principles (1961 : Paris, France) (1971). Statement of principles: adopted at the International Conference on Cataloguing Principles, Paris, October 1961. London: IFLA Committee on Cataloging.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). Functional requirements for bibliographic records final report. Munchen: Saur.

NESSTAR (Networked Social Science Tools and Resources) [Online] Available <http://www.nesstar.org/> [Oct. 11, 2000]

NetLibrary (2000) The MetaText digital textbook solution. [Online] Available <http://www.netlibrary.com/metatext.asp> [Oct. 11, 2000]

Taylor, Arlene G. (1999). The Organization of information. Englewood, Colo.: Libraries Unlimited.

Yale University Library. Imaging America Project. [Online] Available <http://www.library.yale.edu/art/Imaging_America.html> [Oct. 11, 2000]

Crossing a Digital Divide: AACR2 and Unaddressed Problems of Networked Resources

Comments by Glenn Patton

Presented at the Library of Congress Bicentennial Conference on Bibliographic Control in the New Millennium

November 15-17, 2000

Final version

Many of you, I'm sure, remember this famous exchange from the 1967 movie, *The Graduate*, where Mr. McGuire (played by Walter Brooke) completely confuses Ben Braddock (played by Dustin Hoffman) with this cryptic conversation:

> Mr. McGuire: Come with me for a minute. Ben - I just want to say one word to you - just one word.
> Ben Braddock: Yes, sir.
> Mr. McGuire: Are you listening?
> Ben Braddock: Yes, I am.
> Mr. McGuire (gravely): "Plastics."[1]

Inflation being what it has been over the past 30 years, it shouldn't be surprising that there might now be more than one word that I want to share with you without, I hope, similarly confusing you.

Actually, I have three words related to topics in Matthew Beacom's paper that I would like to spend my brief time with this afternoon:

- Publication
- Hierarchy
- Granularity

I want to make some comments about each and ask a few questions in hopes that these may stimulate some discussions during both the general sessions and in the small group discussions.

Publication
Is there a rationale for considering all networked resources published?

Beacom notes in his paper that "some have argued that we should treat everything on the Internet as published." As many of you are aware, that idea first surfaced in Nancy Olson's *Cataloging Internet Resources*,[2] prepared for OCLC as part of two Internet cataloging projects. It was subsequently included in provisions of the *International Standard Bibliographic Description for Electronic Resources* (ISBD(ER)).[3]

From my perspective as someone who participated in the decision to include this in the OCLC guidelines, part of the reason for doing so was the view that, as Beacom notes, "the Internet makes it so easy 'to place materials before the public.'" (one of the dictionary definitions of "publish"). However, there was also a pragmatic aspect to the recommendation. Over the years since high-quality photocopiers and laser printers became prevalent, my OCLC colleagues and I had spent what seemed like an inordinate amount of time helping catalogers define whether "borderline" publications (like genealogies, local histories, other local publications) were really "published." Much of this seemed to fall into the category of "unproductive" dithering that didn't, in the end, make any significant difference in access to the materials. At least in part, the guideline was designed to make that a moot point for similar electronic resources, a pragmatic view that I still share.

What does "publication" mean as we move from "find" to "identify" to "select" to "acquire/obtain access to/use"?

Matthew Beacom raises an excellent question when he asks, "What does 'published' mean on the Internet?" That is a question that we need to consider not only as catalogers but also in relation to other uses of bibliographic data. For example, we all know that "publisher" plays a significant role in the selection process. Think of all those approval plans that bring in all the publications of a particular publisher on a particular subject. They're set up that way because of the reputation of the publisher. Is there a parallel for networked resources?

Hierarchies

Whole <--> part relationships in the networked environment

Moving on to "hierarchies" and linking, Bernhard Eversberg, one of our German colleagues who's participated in the list discussions, has raised the issue of whole/part relationships and how U.S. cataloging practice for many kinds of multi-part items is an impediment to sharing data internationally.[4]

Networked resources offer new possibilities for linking and we need to explore the potential for linking different types of records together, perhaps linking bibliographic descriptions at the collection level to other types of metadata for individual items in that collection.

A shift from "passive" to "active and immediate" hierarchical relationships?

It also seems to me that the shift from relatively passive relationships such as those expressed in print publications by "series title-pages" that give information about other volumes in the series or even lists of works by the same author to much more immediate and "in your face" relationships such as a page where the user is exposed not only to the table of contents for an issue of the electronic journal, *IMF Staff Papers*, but also to basically all of the information that is available at the International Monetary Fund's web site.[5]

Granularity

Are we seeing a return to the 19th-century mixed catalog?

The third word for today is "granularity." Matthew Beacom mentions the possibility of a return to article-level cataloging. In a posting to the BIBCONTROL list, Pauline Cochrane reminds us that, in the 19th century, library catalogs sometimes contained journal article indexing (before we gave all that over to the commercial indexing and abstracting services).[6] As a result, Chapter 13 of AACR2, and its equivalents in previous versions of the rules, are among the least used portions of those rules.

It also seems clear that, in addition to the e-journal aggregations and article databases that seem to be transforming journal publishing, much of what is available on the Internet shares characteristics of "essays in a collection" or "chapters in a larger work" to mention only a couple of other targets for traditional In-analytics.

At what level of granularity are CORC participants creating records?

One thing that has become obvious in working the CORC project participants is the potential need for guidance about what is the appropriate level to describe a networked resource. Do you describe only at the "site" level or at a level below that -- a subsite that forms some kind of logical unit -- or at the individual item level, be that a paper or article, an image, or some other kind of resource?

To aid in looking at this issue, my colleague, Chandra Prabha of the OCLC Office of Research, has been examining a set of CORC resource descriptions created during the period from July 1999 through June 2000. One of the characteristics that she has looked at is the granularity of the cataloging unit. Preliminary analysis indicates that 60% of the resources describe something that appears to be a "whole" item while 33% represent something that is a part of a larger whole with the remaining 7% falling into a gray area that cannot be easily categorized.

This issue is very much involved with the question of "how can we ever hope to control something so vast and changeable as the Web" and I hope that one of the outcomes of this conference might be the beginning of some guidance on the issue of cataloging granularity. I think we all understand the idea that we're not cataloging "every takeout menu and place mat," as Robin Wendler noted in her comments, but catalogers need some help determining what it is that they are or should be cataloging.

A Parting Thought

I ran across a quotation in a recent issue of *The Economist* that made me think about the current state of cataloging for networked resources:

> "Everything that can be invented has been invented." With these sweeping words, the Commissioner of the United States Office of Patents recommended in 1899 that his office be abolished, so spectacular had been the wave of innovation in the late 19th century.[7]

Beacom's 12 recommendations make it clear that "everything that can be cataloged has not been cataloged." Action on these recommendations would give us the consistency and the flexibility to handle networked resources ... and whatever else is lurking around the corner ... and, in the process, keep the library and the cataloger in the center of a networked environment.

---

1. Willingham, Calder, and Buck Henry. *The Graduate*. In *Best American Screen Plays*. First series. New York: Crown Publishers, 1986, p. 300.
2. Olson, Nancy B., ed. *Cataloging Internet Resources: A Manual and Practical Guide*. 2nd ed. Dublin, OH: OCLC, 1997 (http://www.purl.org/oclc/cataloging-internet)
3. *International Standard Bibliographic Description for Electronic Resources: Revised from the ISBD(CF): International Standard Bibliographic Description for Computer Files*, recommended by the ISBD(CF) Review Group. Muenchen: K. G. Saur, 1997. (http://www.ifla.org/VII/s13/pubs/isbd.htm)
4. Eversberg, Bernhard. Beacom's Paper: Posting to BIBCONTROL (forwarded by J. McRee Elrod). October 27, 2000.
5. http://www.imf.org/external/pubs/ft/staffp/1999/09-99/index.htm
6. Cochrane, Pauline. Alternative Architecture: Posting to BIBCONTROL. October 4, 2000.
7. Woodall, Pam. "Untangling E-conomics: a Survey of the New Economy." *The Economist*, 356, no. 8189 (23 Sept. 2000), supplement, p. 5.

Library of Congress
December 21, 2000
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Lois Mai Chan

Professor
School of Library and Information Science
University of Kentucky
Lexington, KY 40506-0039

Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources

About the presenter:

Lois Mai Chan is the author of numerous articles and books on cataloging, classification, subject indexing and online retrieval, including *Library of Congress Subject Headings: Principles and Application*; *Cataloging and Classification: An Introduction*; and *Immroth's Guide to the Library of Congress Classification*. She co-authored *Dewey Decimal Classification: A Practical Guide and Thesauri Used in Online Databases*. From 1986 to 1991, she served as the chair of the Decimal Classification Editorial Policy Committee. Her research interests include classification, controlled vocabulary, authority control, metadata, and retrieval of Web resources. In 1989, she was awarded the Margaret Mann Citation for Outstanding Achievement in Cataloging and Classification given by ALA. In 1992, she received the Distinguished Service Award from the Chinese-America n Librarians Association. In 1999, she was chosen for the Best of LRTS Award for the Best Article Published in 1998.

Full text of paper is available

Summary:

Vocabulary control for improved precision and recall and structured

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Cataloging Directorate Home Page

Library of Congress Home Page

organization for efficient shelf location and browsing have contributed to effective subject access to library materials. The question is whether existing tools can continue to function satisfactorily in dealing with web resources. In our effort to identify library resource description needs and future directions, the online public access catalog (OPAC) should be viewed as a part of the overall information storage and retrieval apparatus on the web rather than something apart from it. Deliberations on the future of bibliographic control and the tools used for its implementation should take into consideration the nature of the web, the characteristics of web resources, and the variety of information retrieval approaches and mechanisms now available and used on the web. Operational conditions on the web are often less structured than in the OPAC environment. While traditional subject access tools such as subject headings and classification schemes have served library users long and well, there are certain limitations to their extended applicability to networked resources. These include the need of trained catalogers for their proper application according to current policies and procedures, the cost of maintenance, and their incompatibility with most tools now used on the web. To meet the challenges of web resources, certain operational requirements must be taken into consideration, the most important being the ability to handle a large volume of resources efficiently and interoperability across different information environments and among a variety of retrieval models. Schemes that are scalable in semantics and flexible in syntax, structure, and application are more likely to be capable of meeting the requirements of a diversity of information retrieval environments and the needs of different user communities. Library of Congress Subject Headings (LCSH), the Library of Congress Classification (LCC), and the Dewey Decimal Classification (DDC) have long been the main staples of subject access tools in library catalogs. Recent deliberations of the Association of Library Collections and Technical Services (ALCTS) Subcommittee on Subject Analysis and Metadata and research findings suggest that in order to extend their usefulness as subject access tools in the web environment, traditional schemes must undergo rigorous scrutiny and re-thinking, particularly in terms of their structure and the way they are applied. Experimentation conducted on subject access schemes in surrogate-based WebPACs and metadata-processed systems demonstrate the potential benefit of structured approaches to description and organization of web resources. Research findings indicate that sophisticated technology can be used to extend the usefulness and to enhance the power of traditional tools. Together, they can provide approaches to content retrieval that may offer improved or perhaps even better subject access than many methods currently used in full-text document analysis and retrieval on the web.

Diane Vizine-Goetz,

commentator

Office of Research
5454 Frantz Rd.
Dublin, OH 43017-3395

About the commentator:

Diane Vizine-Goetz is a research scientist in the Office of Research at OCLC. She joined OCLC in 1983 as a post-doctoral fellow to continue research on database quality she began as a doctoral student. Since then, she has conducted research on the application and use of Library of Congress Subject Headings in online systems and on the development of classifier-assistance tools. She is principal research investigator on a project to enhance the usefulness of the Dewey Decimal Classification as a knowledge organization tool for electronic resources.

Full text of commentary is available

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

Exploiting LCSH, LCC, and DDC To Retrieve Networked Resources Issues and Challenges

Lois Mai Chan
School of Library and Information Science
University of Kentucky
Lexington, KY 40506-0039

Final version

Introduction

The proliferation and the infinite variety of networked resources and their continuing rapid growth present enormous opportunities as well as unprecedented challenges to library and information professionals. The need to integrate Web resources with traditional types of library materials has necessitated a re-examination of the established, well-proven tools that have been used in bibliographic control. Librarians confront new challenges in extending their practice in selecting and organizing library materials to a variety of resources in a dynamic networked environment. In this environment, the tension between quality and quantity has never been keener. Providing quality access to a large quantity of resources poses special challenges.

This paper examines how the nature of the Web and characteristics of networked resources affect subject access and analyses the requirements of effective indexing and retrieval tools. The current and potential uses of existing tools and possible courses for future development will be explored in the context of recent research.

A New Environment and Landscape

For centuries librarians have addressed issues of information storage and retrieval and have developed tools that are effective in handling traditional materials. However, any deliberation on the future of traditional tools should take into consideration the characteristics of networked resources and the nature of information retrieval on the Web. The sheer size demands efficient tools; it is a matter of economy. I will begin by reviewing briefly the nature of the OPAC and the characteristics of traditional library resources. OPACs are by and large homogeneous, at least in terms of content organization and format of presentation, if not in interface design. They are standardized due to the common tools

(AACR2R/MARC, LCSH, LCC, DDC, etc.) used in their construction, and there is a level of consistency among them. The majority of resources represented in the OPACs, i.e., traditional library materials, typically manifest the following characteristics:

- tangible (they represent physical items)
- well-defined (they can be defined and categorized in terms of specific types, such as books, journals, maps, sound recordings, etc.)
- self-contained (they are packaged in recognizable units)
- relatively stable (though subject to physical deterioration, they are not volatile)

The World Wide Web, on the other hand, can be described as vast, distributed, multifarious, machine-driven, dynamic/fluid, and rapidly evolving. Electronic resources, in contrast to traditional library materials, are often:

- amorphous
- ill-defined
- not self-contained
- unstable
- volatile

Over the years, standards and procedures for organizing library materials have been developed and tested. Among these is the convention that trained catalogers and indexers typically carry the full responsibility for providing metadata through cataloging and indexing. In contrast, the networked environment is still developing, meaning that appropriate and efficient methods for resources description and organization are still evolving. Because of the sheer volume of electronic resources, many people without formal training in bibliographic control, including subject specialists, public service personnel, and non-professionals, are now engaged in the preparation and provision of metadata for Web resources. Additionally, the computer has been called on to carry a large share of the labor involved in information processing and organization. The results are often amazing and sometimes dismaying. This raises the question of how to maintain consistency and quality while struggling to achieve efficiency. The answer perhaps lies somewhere between a total reliance on human power and a complete delegation to technology.

Retrieval Models

The new landscape presented by the Web challenges established information retrieval models to provide the power to navigate networked resources with the same levels of efficiency in precision and recall achieved with traditional resources. In her deliberation of subject cataloging in the online environment, Marcia J. Bates pointed out the importance of bringing into consideration search capabilities "*Online search capabilities themselves constitute a form of indexing*. Subject access to online catalogs is thus a combination of original indexing and what we might call 'search capabilities indexing'" (Bates 1989). In contemplating the most effective subject approaches to networked resources, we need to take into account the different models currently used in information retrieval. In addition to the Boolean model, various ranking algorithms and other retrieval models are also implemented. The Boolean model, based on exact matches, is used in most OPACs and many commercial databases. On the other hand, the vector and the probabilistic models are common on the Web, particularly in full-text analysis, indexing, and retrieval (Korfhage 1997; Salton 1994). In these models, the loss of specificity normally expected from traditional subject access tools is compensated to a certain degree by methods of statistical ranking and computational linguistics, based on term occurrences, term frequency, word proximity, and term weighting. These models do not always yield the best results but, combined with automatic methods in text processing and indexing, they have the ability to handle a large amount of data efficiently. They also give some indication of future trends and developments.

Subject Access on the Web

What kinds of subject access tools are needed in this environment? We may begin by defining their functional requirements. Subject access tools are used:

- to assist searchers in identifying the most efficient paths for resource discovery and retrieval
- to help users focus their searches
- to enable optimal recall
- to enable optimal precision
- to assist searchers in developing alternative search strategies
- to provide all of the above in the most efficient, effective, and economical manner

To fulfill these functions in the networked environment, there are certain operational requirements, the most important of these being interoperability and the ability to handle a large amount of resources efficiently. The blurred boundaries of information spaces demand that disparate systems can work together for the benefit of the users. Interoperability enables users to search among resources from a multitude of sources generated and organized according to different standards and approaches. The sheer size of the Web demands attention and presents a particularly critical challenge. For years, a pressing issue facing the libraries has been the large backlogs. If the definition of arrearage is a large number of books waiting in the backroom to be cataloged, then think of Web resources as a huge arrearage sitting in the front yard. How to impose bibliographic control on those resources of value in the most efficient and economical manner possible -- in essence achieving scalability -- is an important mission of the library and information profession. To provide users with a means to seamlessly utilize these vast resources, the operational requirements may be summarized as:

- interoperability among different systems, metadata standards, and languages
- flexibility and adaptability to different information communities, not only different types of libraries, but also other communities such as museums, archives, corporate information systems, etc.
- extensibility and scalability to accommodate the need for different degrees of depth and different subject domains
- simplicity in application, i.e., easy to use and to comprehend
- versatility, i.e., the ability to perform different functions
- amenability to computer application

In 1997, in order to investigate the issues surrounding subject access in the networked environment, ALCTS (Association of Library Collections and Technical Services) established two subcommittees: Subcommittee on Metadata and Subject Analysis and Subcommittee on Metadata and Classification. Their reports are now available (ALCTS 1999, 1999a). Some of their recommendations will be discussed later in this paper.

Verbal Subject Access

While subject access to networked resources is available, there is much room for improvement. Greater usage of controlled vocabulary may be one of the answers. During the past three decades, the introduction and increasing popularity and, in some cases, total reliance on free-text or natural language searching have brought a key question to the forefront: Is there still a need for controlled vocabulary? To information professionals who have appreciated the power of controlled vocabulary, the answer has always been a confident "yes." To others, the affirmative answer became clear only when searching began to be bogged down in the sheer size of retrieved results. Controlled vocabulary offers the benefits of consistency, accuracy, and control (Bates 1989), which are often lacking in the free-text approach. Even in the age of automatic indexing and with the ease in keyword searching, controlled vocabulary has much to offer in improving retrieval results and in alleviating the burden of synonym and homograph control placed on the user. For many years, Elaine Svenonius has argued that using controlled vocabulary retrieves more relevant records by placing the burden on the indexer rather than the user (Svenonius 1986; Svenonius 2000). Recently, David Batty makes a similar observation on the role of controlled vocabulary in the Web environment: "There is a burden of effort in information storage and retrieval that may be shifted from shoulder to shoulder, from author, to indexer, to index language designer, to searcher, to user. It may even be shared in different proportions. But it will not go away" (Batty 1998).

Controlled vocabulary most likely will not replace keyword searching, but it can be used to supplement and complement keyword searching to enhance retrieval results. The basic functions of controlled vocabulary, i.e., better recall through synonym control and term relationships and greater precision through homograph control, have not been completely supplanted by keyword searching, even with all the power a totally machine-driven system can bring to bear. To this effect, the ALCTS Subcommittee on Metadata and Subject Analysis recommends the use of a combination of keywords and controlled vocabulary in metadata records for Web resources (ALCTS 1999a).

Subject heading lists and thesauri began as catalogers' and indexers' tools, as a source of, and an aid in choosing, appropriate index terms. Later, they were also made available to users as searching aids, particularly in online systems. Traditionally, controlled vocabulary terms embedded in metadata records have been used as a means of matching the user's information needs against the document collection. Subject headings and descriptors, with their attendant equivalent and related terms, facilitate the searcher's ability to make an exact match of search terms against assigned index terms. Manual mapping of users' input terms to controlled vocabulary terms--for example, consulting a thesaurus to identify appropriate search terms--is a tedious process and has never been widely embraced by end-users. With the availability of online thesaurus-display, the mapping is greatly facilitated by allowing the user to browse and select controlled vocabulary terms in searching. Controlled vocabulary thus serves as the bridge between the searcher's language and the author's language.

Even in free-text and full-text searching, keywords can be supplemented with terms "borrowed" from a controlled vocabulary to improve retrieval performance. Participating researchers of TREC (the Text Retrieval Conference), the large-scale cross-system search engine evaluation project, have found that "the amount of improvement in recall and precision which we could attribute to NLP [natural language processing] appeared to be related to the type and length of the initial search request. Longer, more detailed topic statements responded well to LMI [linguistically motivated indexing], while terse one-sentence search directives showed little improvement" (Strzalkowski et al. 2000). Because of the term relationships built in a controlled vocabulary, the retrieval system can be programmed to automatically expand an original search query to include equivalent terms, post-up or down to hierarchically related terms, or suggest associative terms. Users typically enter simple natural language terms (Drabenstott 2000), which may or may not match the language used by authors. When the searcher's keywords are mapped to a controlled vocabulary, the power of synonym and homograph control could be invoked and the variants of the searcher's terms could be called up (Bates 1998). Furthermore, the built-in related controlled terms could also be brought up to suggest alternative search terms and to help users focus their searches more effectively. In this sense, controlled vocabulary is used as a query-expansion device. It can be used to complement uncontrolled terms and terms from lexicons, dictionaries, gazetteers, and similar tools, which are rich in synonyms, but often lacking in relational terms. In the vector and probabilistic retrieval models, using a conflation of variant and related terms often yield better results than relying on the few "key" words entered by the searcher. Equivalent and related terms in a query provide context for each other. Including additional search terms from a controlled vocabulary can improve the ranking of retrieved items.

Classification and Subject Categorization

With regard to knowledge organization, traditionally, classification has been used in American libraries primarily as an organizational device for shelf-location and for browsing in the stacks. It has often been used also as a tool for collection management, for example, assisting in the creation of branch libraries and in the generation of discipline-specific acquisitions or holdings lists. In the OPAC, classification has regained its bibliographic function through the use of class numbers as access points to MARC records. To continue the use of class numbers as access points, the ALCTS Subcommittee on Metadata and Subject Analysis recommends that this function be extended to other types of metadata records by including in them class numbers, but not necessarily the item numbers, from existing classification schemes (ALCTS 1999a).

In addition to the access function, the role of classification has been expanded to those of subject browsing and navigational tools for retrieval on the Web. In its study of the use of classification devices for organizing metadata, the ALCTS Subcommittee on Metadata and Classification has identified seven functions of classification: location, browsing, hierarchical movement, retrieval, identification, limiting/partitioning, and profiling (ALCTS 1999).

With the rapid growth of networked resources, the enormous amount of information available on the Web cries out for organization. When subject categorization devices first became popular among Web information providers, they resembled broad classification schemes, but many were lacking the rigorous hierarchical structure and careful conceptual organization found in established schemes. Many library portals, which began with a collection of a limited number of selected electronic resources offering only keyword searching and/or an alphabetical listing, have adopted broad subject categorization schemes when the collection of electronic resources became voluminous and unwieldy (Waldhart et al. 2000). Some of these subject categorization devices are based on existing classification schemes, e.g., Internet Public Library Online Texts Collection based on the Dewey Decimal Classification (DDC) and CyberStacks(sm) based on the Library of Congress Classification (LCC) (McKiernan 2000); others represent home-made varieties.

Subject categorization defines narrower domains within which term searching can be carried out more efficiently and enables the retrieval of more relevant results. Combination of subject categorization with term searching has proven to be an effective and efficient approach in resource discovery and data mining. In this regard, classification or subject categorizing schemes function as information filters, used to efficiently exclude large segments of a database from consideration of a search query (Korfhage 1997).

Recent Research on Subject Access Systems

Before we explore the potential directions for future development of traditional subject access tools, let us also examine some of the recent research efforts and their implications for current and future methods of subject indexing and access. A huge body of research has been reported in the literature. Three areas of experimentation that I consider to have important bearings on subject access tools are automatic indexing, mapping terms and data from different sources, and integrating different subject access tools.

Automatic indexing

In the past few decades, some of the most important research in the field of information storage and retrieval has been focused on automatic indexing. Beginning with the pioneer efforts in the 1970s, various techniques, including term weighting, statistical analysis of text, and computational linguistics, have been developed and applied. More recent examples include OCLC's Scorpion project, which uses automatic methods to perform subject recognition and to generate machine-assigned DDC numbers for electronic resources (Shafer 1997). Another OCLC project, WordSmith, (Godby and Reighart 1998), applying computational linguistics to implement a series of largely statistical filters, investigates the feasibility of extracting subject terminology directly from raw text. An extension of this project, called Extended WordSmith, applies a similar technique to the automatic generation of thesaural terms. On the more practical side, the recent implementation of the LEXIS/NEXIS SmartIndexing Technology combines controlled vocabulary features with an indexing algorithm to arrive at a relevance score or percentage based on criteria such as term frequency, weight, and location in document in indexing LEXIS/NEXIS news collections (Quint 1999; Tenopir 1999).

Mapping terms and data from different sources

Mapping natural-language expressions typical of end-user search queries and of automatically extracted index terms to more structured subject-language is an area that has been explored and holds great promise (Svenonius 2000). A recent example is the "Entry Vocabulary Modules" project at the University of California-Berkeley, which explores the possibility of mapping "ordinary language queries" to indexing terms based on metadata subject vocabularies unfamiliar to the user, including classification numbers, subject headings, and descriptors from various subject- or domain-specific vocabularies (Buckland et al. 1999).

On another front, numerous efforts have focused on mapping subject data from different vocabulary sources, including free-text terms extracted from full texts, controlled vocabularies, classification data, and name authority data. Because the networked environment is open and multifarious, multiple tools for resource description and subject access are often used side-by-side. In this open environment, use of multiple controlled vocabularies within the same system is not uncommon. Harmonization of different vocabularies, similar or analogous to crosswalks among metadata schemes, is an important issue. Even before the advent of the World Wide Web, mapping subject terms from multiple thesauri was a topic of great interest and concern. An example was Carol Mandel's investigation to resolve the problems caused by using multiple vocabularies within the same online system (Mandel 1987). Much progress has been made in biomedical vocabularies. The Unified Medical Language System (UMLS) Metathesaurus currently maps biomedical terms from over fifty different biomedical vocabularies, some in multiple languages (Nelson 1999; National Library of Medicine 2000). A general metathesaurus covering all subjects is still lacking. Outside of the library context, there are also efforts to map index terms from different sources. An example is WILSONLINE's OmniFile, which results from merging index terms from six H.W. Wilson indexes into one index file.

On a broader scale, indexes from different language sources also need to be interoperable. Mapping between controlled vocabularies in different languages is an issue of great interest particularly in the international community. MACS (Multilingual ACcess to Subject), an ongoing international project involving Swiss, German, French, and British national libraries, attempts to link subject authority files in three different languages, Schlagwortnormdatei (SWD, German), RAMEAU (French), and the Library of Congress Subject Headings (English) (Landry 2000).

Mapping between subject headings and class numbers is not new. Past efforts have focused mainly on facilitating subject cataloging and indexing. Examples include the linking of many LCC numbers to headings in the *Library of Congress Subject Headings (LCSH)* list and the inclusion of abridged DDC numbers in the *Sears List of Subject Headings* (Sears). More recently, there have been efforts to map between DDC numbers and LCSH (Vizine-Goetz 1998). OCLC's WordSmith project mentioned earlier demonstrates that subject terms can be identified and extracted automatically from raw texts and mapped to existing classification schemes such as DDC (Godby and Reighart 1998). Diane Vizine-Goetz demonstrates how results from the research projects WordSmith and ExTended Concept Trees can be used together to enhance DDC (Vizine-Goetz 1997). The same techniques should be applicable to LCC also. With the implementation of the CORC (Cooperative Online Resource Catalog) project, results of many of OCLC's research projects have converged in practice. Actual application includes the automatic generation of subject data and DDC numbers in metadata records. A most impressive feature of CORC that can yield great benefit is the capability of mapping names and subject words and phrases input by catalogers or indexers or those automatically generated from websites to entries in subject and name authority files.

Integrating different subject access tools

In the manual environment, subject headings and classification systems have more or less operated in isolation from each other. Technology offers the possibility of integrating tools of different sorts to enhance retrieval results as well as facilitate subject cataloging and indexing. The merging or integration of classification with controlled vocabulary holds great potential. Numerous research projects have been undertaken and some of the designs have been tested. For example, Karen Markey's project incorporated the Dewey Decimal Classification as a retrieval tool alongside subject searching in an online system (Markey 1986). Her research was built on AUDACIOUS, an earlier project using UDC as the index language with nuclear science literature (Freeman and Atherton 1968).

In a system called Cheshire, Ray Larson, used a method called "classification clustering," combined with probabilistic retrieval techniques, to improve subject searching in the OPAC. Starting with LC call numbers and using probabilistic ranking and weighting mechanisms, Larson demonstrates that class numbers combined with subject terms generated from titles of documents and subject headings in MARC records can enhance access points and improve greatly the retrieval results. The integration of different types of access is significant, as Larson observes: "The topical access points of the MARC records used in online catalogs, such as the classification numbers, subject headings, and title keywords, have usually been treated in strict isolation from each other in search. The classification clustering method is one way of effectively combining these difference 'clues' to the database contents" (Larson 1991).

Traditional Tools in the Networked Environment

The concepts surrounding subject access have been explored in relation to the configuration of the Web landscape and retrieval models. In this context, a question that can be raised is: How well can existing subject access tools fulfill the requirements of networked resources? More specifically, how adequate are traditional tools such as LCSH, LCC, and DDC in meeting the challenges of effective and efficient subject retrieval in the networked environment?

Library of Congress Subject Headings

With regard to LCSH specifically, a basic question is whether a new controlled vocabulary more suited to the requirements of electronic resources should be constructed. The ALCTS Subcommittee on Metadata and Subject Analysis deliberated on this question and examined the options relating to the choice of subject vocabulary in metadata records. After considering the options of developing a new vocabulary or adopting or adapting one or more existing vocabularies, the Subcommittee recommends the latter option (ALCTS 1999a). For a general controlled vocabulary covering all subjects, the Subcommittee recommends the use of LCSH or Sears with or without modifications. Among the reasons for retaining

LCSH are: (1) LCSH is a rich vocabulary covering all subject areas, easily the largest general indexing vocabulary in the English language; (2) there is synonym and homograph control; (3) it contains rich links (cross references indicating relationships) among terms; (4) it is a pre-coordinate system that ensures precision in retrieval; (5) it facilitates browsing of multiple-concept or multi-faceted subjects; and, (6) having been translated or adapted as a model for developing subject headings systems by many countries around the world, LCSH is a de facto universal controlled vocabulary. In addition, there is another major advantage. Retaining LCSH as subject data in metadata records would ensure semantic interoperability between the enormous store of MARC records and metadata records prepared according to various other standards.

While the vocabulary, or semantics, of LCSH has much to contribute to the management and retrieval of networked resources, the way it is currently applied has certain limitations: (1) because of its complex syntax and application rules, assigning LC subject headings according to current Library of Congress policies requires trained personnel; (2) subject heading strings in bibliographic or metadata records are costly to maintain; (3) LCSH, in its present form and application, is not compatible in syntax with most other controlled vocabularies; and, (4) it is not amenable to search engines outside of the OPAC environment, particularly current Web search engines. These limitations mean that applying LCSH properly in compliance with current policy and procedures entails the following requirements:

- trained catalogers and indexers
- systems with index browsing capability
- systems with online thesaurus display
- sophisticated users (Drabenstott 1999)

In the networked environment, such conditions often do not prevail. What direction and steps need to be taken for LCSH to overcome these limitations and remain useful in its traditional roles as well as to accommodate other uses? Pondering the viability of LCSH in the networked environment, the ALCTS Subcommittee on Metadata and Subject Analysis recommends separating the consideration regarding semantics from that relating to application syntax, in other words, distinguishing between the vocabulary (LCSH per se) and the indexing system (i.e., how LCSH is applied in a particular implementation).

This recommendation involves several important concepts that need to be reviewed. Semantics and syntax are two distinct aspects of a controlled vocabulary. Semantics concerns the source vocabulary, i.e., what appears in the term list (e.g., a thesaurus or a subject headings list) that contains the building blocks for constructing indexing terms or search statements. It covers the scope and depth, the selection of terms to be included, the forms of valid terms, synonym and homograph control, and the syndetic (cross-referencing) devices. Semantics should be governed by well-defined principles of vocabulary structure.

At the heart of the syntax concept is the representation of complex subjects through combination, or coordination, of terms representing different subjects or different facets (defined as families of concepts that share a common characteristic (Batty 1998)) of a subject. There are two aspects of syntax: term construction and application syntax. Term construction, i.e., how words are put together to represent concepts in the thesaurus, is an aspect of semantics and is a matter of principle; while application syntax, i.e., how thesaurual terms are put together to reflect the contents of documents in the metadata record, is a matter of policy, determined by practical factors such as user needs, available resources, and search engines and their capabilities.

Enumeration (i.e., the listing of pre-established multiple-concept index terms in the thesaurus) and faceting (i.e., the separate listing of single-concept or single-facet terms defined in distinctive categories based on common, shared characteristics) are aspects of term construction, while precoordination and postcoordination relate to application syntax. Term combination can occur at any of three stages in the process of information storage and retrieval: (1) during vocabulary construction; (2) at the stage of cataloging or indexing; or, (3) at the point of retrieval. When words or phrases representing different subjects or different facets of a subject are pre-combined at the point of thesaurus construction, we refer to the process as enumeration. When term combination occurs at the stage of indexing or cataloging, we refer to the practice as precoordination. In contrast, postcoordination refers to the combination of terms at the point of retrieval. A totally enumerative vocabulary is by definition precoordinated. On the other hand, a faceted controlled vocabulary--i.e., a system that provides individual terms in clearly defined categories, or facets-- may be applied either precoordinately or postcoordinately. A faceted scheme hence is more flexible. An example of a rigorously faceted, precoordinate system is PRECIS (previously used in the British National Bibliography). Another example is the Universal Subject Environment (USE) system, proposed in a recent article by William E. Studwell, which contains faceted terms and uses special punctuation marks as facet indicators (Studwell 2000). On the other hand, current indexing systems used in abstracting and indexing services employing controlled vocabularies are typically postcoordinated. Whether a precoordinate approach or a postcoordinate approach is used in a particular implementation is a matter of policy and is agency-specific. In the remainder of this paper, we will focus on the semantics and term construction issues.

Because of the varied approaches to retrieval in different search environments and the different needs of diverse user communities, a vocabulary that is flexible enough to be used either precoordinately or postcoordinately would be the most viable. A faceted scheme can accommodate different application syntaxes, from the most complex (e.g., full-string approach typically found in OPACs) to the simplest (descriptor-like terms used in most indexes) and would also allow different degrees of sophistication. The advantages of a faceted controlled vocabulary can be summarized as follows:

- simple in structure
- flexible in application (i.e., able to accommodate a tiered approach to allow different levels of subject representation)
- amenable to software applications (Batty 1998)
- amenable to computer-assisted indexing and validation
- interoperable with the majority of modern indexing vocabularies
- easier and more economical to maintain than an enumerated vocabulary

On the last point regarding efficient thesaurus maintenance, Batty remarks: "Facet procedure has many advantages. By organizing the terms into smaller, related groups, each group of terms can be examined more easily and efficiently for consistency, order, hierarchical relationships, relationships to other groups, and the acceptability of the language used in the terms. The faceted approach is also useful for its flexibility in dealing with the addition of new terms and new relationships. Because each facet can stand alone, changes can usually be made easily in a facet at any time without disturbing the rest of the thesaurus" (Batty 1998). Thus, a faceted LCSH will be easier to maintain. With the current LCSH, updating terminology sometimes can be a tedious operation. For example, when the heading "Moving-pictures" was replaced in 1987 by "Motion pictures," approximately 400 authority records were affected! (El-Hoshy 1998).

A faceted LCSH is by no means a new idea. Earlier advocates of such an approach include Pauline A. Cochrane (1986) and Mary Dykstra (1988). To remain viable in the networked environment, a controlled vocabulary, such as LCSH, must be able to accommodate different retrieval models mentioned earlier as well as different application policies. Outside of the OPAC, most search engines, including many used in library portals for Web resources, lack the ability to accommodate full-string browsing and searching. Even among systems that can handle full strings, their capabilities and degrees of sophistication also vary. With a faceted vocabulary, it will not be an either/or proposition between the precoordinate full-string application and the postcoordinate approach, but rather a question of how LCSH can be made to accommodate both and any variations in between, thus ensuring maximum flexibility and scalability in terms of application. Mechanisms for full-string implementation of LCSH are already in place; for example, in the OPAC environment, with highly trained personnel and the searching and browsing capabilities of integrated systems, the full-string syntax has long been employed in creating subject headings in MARC records. In the heterogeneous environment outside of the OPAC, we need a more flexible system in order to accommodate different applications. LCSH can become such a tool, and its use can be extended to various metadata standards and with different encoding schemes. Investigations and experiments on the viability of LCSH have already begun. Using LCSH as the source vocabulary, FAST (Faceted Application of Subject Terminology), a current OCLC research project, explores the possibility and feasibility of a postcoordinate approach by separating time, space, and form data from the subject heading string (Chan et al. in press).

Now we come to the question of where LCSH stands currently in becoming a viable system for the networked environment. LCSH began in the late nineteenth century as an enumerative scheme. It gradually took on some of the features of a faceted system, particularly in the adoption of commonly used form subdivisions and the increasing use of geographic subdivisions. In the latter part of the twentieth century LCSH has taken further steps, ever so cautiously, in the direction of more rigorous faceting. In 1974, the Library of Congress took a giant leap forward in expanding the application of commonly used subdivisions by designating a large number of frequently used topical and form subdivisions "free-floating," thus allowing great flexibility in application. The adoption of BT, NT, RT in the 11th (1988) edition rendered LCSH more in line with thesaural practice. After the Subject Subdivisions Conference held in 1991 (The Future of Subdivisions, 1992), the Library of Congress has embarked on a program to convert many of the topical subdivisions into topical main headings. Finally, in 1999, the implementation of subfield $v for form subdivisions in the 6xx (subject-related) fields in the MARC format, marking the distinction between form and topical subdivisions, moved LCSH yet another step closer to becoming a faceted system. Considering the gradual steps the Library of Congress has taken over the years, even a person not familiar with the history of LCSH must conclude logically that LCSH is heading in the direction of becoming a fully faceted vocabulary. It is not there yet; but, with further effort, LCSH can become a versatile system that is capable of functioning in heterogeneous environments and can serve as the unified basis for supporting diversified uses while maintaining semantic interoperability among them.

A faceted LCSH has a number of potential uses in the areas of thesaurus development and management, indexing, and retrieval. As mentioned earlier, to enhance the interoperability of a multitude of controlled vocabularies, a general metathesaurus covering all subjects would be most desirable (ALCTS 1999a). It will not be a trivial task, but the first question the library and information profession must agree upon is whether it is something worth pursuing. LCSH, with its rich vocabulary--the largest in the English language--can serve as a basis or core of such a metathesaurus.

From a different perspective, LCSH could also be used as the basis for generating subject- or discipline-specific controlled vocabularies or special-purpose thesauri. The *AC Subject Headings* (formerly *Subject Headings for Children's Literature*) sets a precedent. Other examples include a large "superthesaurus" proposed by Bates (1989), with a rich entry vocabulary as a part of a friendly front-end user interface for the OPAC. While many subject domains and disciplines such as engineering, art, and biomedical sciences have their own controlled vocabularies, many specialized areas and non-library institutions still lack them. These include for-profit as well as non-profit organizations, government agencies, historical societies, special-purpose museums, consulting firms, fashion design companies, to name a few. Many of these rely on their curators or researchers, most of whom have not been trained in bibliographic control, to take responsibility for organizing Internet resources. Having a comprehensive subject access vocabulary to draw and build upon would be of tremendous help in developing their specialized thesauri.

To move LCSH further along the way towards becoming a faceted vocabulary, if indeed such is the direction to be followed, more can be done to its semantics. Aspects of particular concern that need close scrutiny and re-thinking include principles of term selection, enhanced entry vocabulary, rigorous term relationships, and particularly term construction.

Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC)

In recent years, with the support of the OCLC Research Office, DDC has made great strides in adapting to the networked environment and becoming a useful tool for organizing electronic resources. For example, the newly developed WebDewey contains, in addition to the DDC/LCSH mapping feature first developed in Dewey for Windows, an automated classification tool for generating candidate DDC numbers during metadata record creation. It has taken LCC somewhat longer because its voluminous schedules have only recently been converted to the MARC format. Let us hope that the Library Congress can now turn its attention to making LCC a useful tool not only in the library stacks but also as an organizing tool of networked resources. Results and insights gained from experimental and actual implementations of Web application of DDC and other classification schemes should be applicable to LCC as well.

Existing classification schemes have already been adopted or adapted to a limited extent for use as subject categorization devices for Web resources. Examples include the adaptation of DDC in NetFirst and CyberDewey and the use of LCC outlines in Cyberstacks. In this particular role, existing classification schemes need greater flexibility and more attention to their structure. Adaptability of classification schemes can take the form of flexibility in the depth of hierarchy and variability in the collocation of items in the array. The requirement of depth varies from application to application. As a tool for shelf-location and bibliographic arrangement, considerable depth in classification is required, as evidenced in the growth of both DDC and LCC in the past. As a navigating tool typified by the subject categorization schemes used in the popular Web directories, broad schemes are often sufficient. What is needed is a flexibility of depth and the amenability to the creation of classificatory structures focused on specific subject domains. Flexibility in depth has always been a feature of DDC and UDC, with the availability of abridged, medium, and full versions, in recognition of the different needs of school, public, and research libraries. LCC has not yet demonstrated this flexibility. This is an area worth exploring.

The principle of literary warrant, i.e., basing the development of a scheme or system on the nature and extent of resources being described and organized, operates in the Web environment as well as in the print environment. In the development of subject categorization schemes used in popular Web information services, such as Yahoo! and Northern Light (Ward 1999) as well as many library portals, we have often witnessed the gradual extension from simple, skeletal outlines to increasingly elaborate structures--almost a mirror of the development of classification schemes in the early days. Flexibility in the collocation of topics in an array would also be helpful, if the same topics in an array could be arranged or re-arranged in different orders depending on the target audiences. For example, the categorization scheme in NetFirst uses the DDC structure, but modifies the arrangement of the categories to suit its target users (Vizine-Goetz 1997).

Observing recent uses of classification-like structures on the Web and the tortuous re-inventing and re-discovering of classification principles in both research and practice (Soergel 1999a), one sees a need for both broad/general (covering all subjects) and close/detailed (subject- or domain-specific) classification schemes. Portals found on websites of general libraries, ranging from school and public libraries to large academic libraries that cover a broad range of subject domains, need schemes of varying depths with a top-down approach, beginning with the broadest level and moving down to narrower subjects as needed. On the other hand, portals that serve special clientele often need specialized schemes with more details. These often require a bottom-up approach starting with topics identified from a collection of documents focusing on a specific theme or mission. How to organize these topics into a coherent structure has often stymied those not trained in the principles and techniques of knowledge organization. The library and information profession can make a contribution here. Subject taxonomy schemes built around specific disciplines (art, education, human environmental sciences, mathematics, engineering), industries (petroleum, manufacturing, entertainment), consumer-oriented topics (automobiles, travel, sports), and problems (environment, aging, juvenile delinquency) can serve diverse user communities, from special libraries to corporate or industry information centers to personal resource collections.

For domain- and subject-specific organizing schemes I suggest a modular approach. In building special-purpose thesauri mentioned earlier, LCSH could serve as the source vocabulary, and DDC or LCC could be used to facilitate the identification and extraction of terms related to specific subjects or domains and could provide the underlying hierarchical structure. Where more details are needed in a particular scheme, terms can be added to the basic structure as needed, thus making the specialized scheme an extension of the main structure and vocabulary. Developing these modules with a view of fitting them as nodes, even on a very broad level, into the overall classification structures of meta-schemes such as DDC and LCC can go a long way to ensure their future interoperability.

As mentioned earlier, the merging or integration of controlled subject vocabulary with classification in order to facilitate both information storage and retrieval has great potential, because they complement each other. A subject heading or descriptor represents a particular topic treated from all perspectives, while classification gathers related topics viewed from the same perspective. Traditionally, each performs its specific function and contributes to information organization and retrieval more or less in isolation. Together, they have the potential of improving efficiency as well as effectiveness. Schemes simple and logical in design lend themselves to interoperate efficiently with each other. How to combine the salient features of a rich vocabulary like LCSH and the structured hierarchy found in classification schemes such as LCC and DDC to improve retrieval of networked resources remains a fertile field for research and exploration.

Conclusion

The sheer volume of available networked resources demands efficiency in knowledge management. Of course, we intend to provide quality and to maintain consistency also. Content representation schemes and systems design must meet halfway--a combination of the intellect and technology, capitalizing on the power of the human mind and the capabilities of the machine. Technology has provided an impetus in the creation of an enormous amount of information; it can also help in its effective and efficient management and retrieval (Soergel 1999). A proper balance in the distribution of efforts between human intellect and technology can ensure both quality and efficiency in helping users gain the maximum benefits from the rich resources that are available in the networked environment. Already, technology has helped create many useful devices for efficient management and application of traditional tools, for example, Dewey for Windows, the WebDewey, and ClassificationPlus. These developments are encouraging. In the near future, we may expect also new applications which can help us not only do the same things better and more efficiently, but also maximize the power of existing subject access tools hitherto not yet exploited.

References

ALCTS/CCS/SAC/Subcommittee on Metadata and Classification. (1999). Final Report. http://www.ala.org/alcts/organization/ccs/sac/metaclassfinal.pdf

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. (1999a). Subject Data in the Metadata Record: Recommendations and Rationale: A Report from the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. http://www.ala.org/alcts/organization/ccs/sac/metarept2.html

Bates, Marcia J. (1998). Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors. *Journal of the American Society for Information Science.* 49(13):1185-1205.

Bates, Marcia J. (October 1999). The Invisible Substrate of Information Science. *Journal of the American Society for Information Science.* 50(12):1043-1050.

Bates, Marcia J. (October 1989). Rethinking Subject Cataloging in the Online Environment. *Library Resources & Technical Services.* 33(4):400-412.

Bates, Marcia J. (1986). Subject Access in Online Catalogs: A Design Model. *Journal of the American Society for Information Science.* 37:357-76.

Batty, David. (November 1998). WWW -- Wealth, Weariness or Waste: Controlled Vocabulary and Thesauri in Support of Online Information Access. *D-Lib Magazine.* http://www.dlib.org/dlib/november98/11batty.html

Buckland, Michael, et al.. (January 1999). Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. *D-Lib Magazine.* http://www.dlib.org/dlib/january99/buckland/01buckland.html

Burton, Paul F. (Summer 1998). Issues and Challenges of Subject Access. *Catalogue & Index.* 128:1-7.

Chan, Lois Mai, Eric Childress, Rebecca Dean, Edward T. O'Neill, and Diane Vizine-Goetz. (in press). A Faceted Approach to Subject Data in the Dublin Core Metadata Record. *Journal of Internet Cataloging.*

Chan, Lois Mai. (1995). *Library of Congress Subject Headings: Principles and Application*. 3rd ed. Englewood, CO: Libraries Unlimited.

Cochrane, Pauline A. (1986). *Improving LCSH for Use in Online Catalogs: Exercises for Self-Help with a Selection of Background Readings*. Littleton, CO: Libraries Unlimited.

Drabenstott, Karen M. (2000). Web Search Strategies. In *Saving the User's Time through Subject Access Innovation*, edited by William J. Wheeler. Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

Drabenstott, Karen M., Schelle Simcox, and Marie Williams. (Summer 1999). Do Librarians Understand the Subject Headings in Library Catalogs? *Reference & User Services Quarterly*. 38(4):369-87.

Dykstra, Mary. (March 1 1988). LC Subject Headings Disguised as a Thesaurus. *Library Journal*. 113:42-46.

El-Hoshy, Lynn M. (August 1998). Charting a Changing Language with LCSH. *Library of Congress Information Bulletin*. 57(8):201.

Freeman, Robert R. and Atherton, Pauline. (April 1968). *AUDACIOUS - An Experiment with an On-Line, Interactive Reference Retrieval System Using the Universal Decimal Classification as the Index Language in the Field of Nuclear Science.* New York, NY: American Inst. of Physics, (QPX02169).

The Future of Subdivisions in the Library of Congress Subject Headings System: Report from the Subject Subdivisions Conference, sponsored by the Library of Congress, May 9-12, 1991, ed. Martha O'Hara Conway. Washington, DC: Library of Congress, Cataloging Distribution Service, 1992.

Godby, C. J. (1998). The Wordsmith Toolkit. *Annual Review of OCLC Research 1997*. Available at http://www.oclc.org/oclc/research/publications/review97/godby/godby_wordsmith.htm

Godby, C. Jean and Reighart, Ray R. (1998). The WordSmith Indexing System. http://www.oclc.org/oclc/research/publications/review98/godby_reighart/wordsmith.htm

Godby, C. J. and R. Reighart. (1998a). The Wordsmith Project Bridges the Gap between Tokenizing and Indexing. *OCLC Newsletter*, July 1998. Available at http://www.oclc.org/oclc/new/n234/rsch_wordsmith_research_project.htm.

Korfhage, Robert R. (1997). The Matching Process. In Information Storage and Retrieval. New York: Wiley. (pp. 79-104).

Landry, Patrice. (2000). The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD). Classification and Indexing Workshop, 66th IFLA Council and General Conference, Meeting No. 181. http://www.ifla.org/IV/ifla66/papers/165-181e.pdf

Larson, Ray R. (1991). Classification Clustering, Probabilistic Information Retrieval, and the Online Catalog. *The Library Quarterly* 61(2):133-173.

Mandel, Carol A. (1987). *Multiple Thesauri in Online Library Bibliographic Systems: A Report Prepared for Library of Congress Processing Services.* Washington, DC: Library of Congress, Cataloging Distribution Service.

Markey, Karen and Anh N. Demeyer. (1986). *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog: Final Report to the Council on Library Resources.* Dublin, OH: OCLC.

Moen, William E. (March, 2000). Interoperability for Information Access: Technical Standards and Policy Considerations. *Journal of Academic Librarianship*. 26(2):129-32.

National Library of Medicine. (February 2000). Fact Sheet: UMLS (r) Metathesaurus (r) http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Nelson, Stuart J. (1999). The Role of the Unified Medical Language System (UMLS) in Vocabulary Control: CENDI Conference "Controlled Vocabulary and the Internet." http://www.dtic.mil/cendi/pres_arc.html

Olson, Tony and Gary Strawn. (March 1997). Mapping the LCSH and MeSH Systems. *Information Technology and Libraries*. 16(1):5-19.

Quint, Barbara. (December 1999). Company Dossier Product Emerges from LEXIS-NEXIS SmartIndexing Technology. *Information Today*. 16(1):18-19.

Salton, Gerard. (February 1994). Automatic Structuring and Retrieval of Large Text Files. *Communications of the ACM*. 37:97-108.

Salton, Gerard. (1991). Developments in Automatic Text Retrieval. *Science*. 253:974-980.

Shafer, Keith. (October/November 1997). Scorpion Helps Catalog the Web. *Bulletin of the American Society for Information Science*. 24(1):28-29.

Shafer, Keith E. (1996). Automatic Subject Assignment Via the Scorpion System. *Annual Review of OCLC Research*, pp. 20-21.

Soergel, Dagobert. (September 1999). Enriched Thesauri as Networked Knowledge Bases for People and Machines: Paper Presented at the CENDI Conference Controlled Vocabulary and the Internet, Bathesda, MD, 1999 September 29. http://www.dtic.mil/cendi/presentations/cendisoergel.pdf

Soergel, Dagobert. (1999a). The Rise of Ontologies or the Reinvention of Classification. *Journal of the American Society for information Science*. 50(12):1119-1120.

Strzalkowski, Tomek et al.. (2000). Natural Language Information Retrieval: TREC-8 Report. http://trec.nist.gov/pubs/trec8/papers/ge8adhoc2.pdf

Studwell, William E. (2000). USE, the Universal Subject Environment: A New Subject Access Approach in the Time of the Internet. *Journal of Internet Cataloging*. 2(3/4):197-209.

Svenonius, Elaine. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, MA: MIT Press.

Svenonius, Elaine. (1986). Unanswered Questions in the Design of Controlled Vocabularies. *Journal of the American Society for Information Science*. 37:331-40.

Tenopir, Carol. (November 1, 1999). Human or Automated, Indexing Is Important. *Library Journal*.

124(18):34,38.

Vizine-Goetz, Diane. (1996). Classification Research at OCLC. *Annual Review of OCLC Research*, pp. 27-33.

Vizine-Goetz, Diane. (October/November 1997). From Book Classification to Knowledge Organization: Improving Internet Resource Description and Discovery. *Bulletin of the American Society for Information Science.* 24(1):24-27.

Vizine-Goetz, Diane. (May/June 1998). Subject Headings for Everyone: Popular Library of Congress Subject Headings with Dewey Numbers. *OCLC Newsletter*. 233:29-33.

Waldhart, Thomas J., Joseph B. Miller, and Lois Mai Chan. (March 2000). Provision of Local Assisted Access to Selected Internet Information Resources by ARL Academic Libraries. *Journal of Academic Librarianship*. 26(2):100-109.

Ward, Joyce. (1999). Indexing and Classification at Northern Light: Presentation to CENDI Conference "Controlled Vocabulary and the Internet," Sept 29, 1999. http://www.dtic.mil/cendi/pres_arc.html

Younger, Jennifer A. (Winter 1997). Resource Description in the Digital Age. *Library Trends*. 45(3):462-87.

Library of Congress
December 19, 2000
Comments: lcweb@loc.gov

Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources

Comments by Diane Vizine-Goetz

Final version

Conference goal

The conference goal for the Library of Congress Subject Headings (LCSH), Library of Congress Classification (LCC) and the Dewey Decimal Classification (DDC) is to encourage wider use of these schemes for resource description and discovery. In considering new uses for our traditional subject access systems, it is useful to review how widely these schemes are currently used. Sometimes we forget the role they play in subject retrieval worldwide as we are overwhelmed by news of what the web is doing or not doing.

As reported by Magda Heiner-Freiling, we find that LCSH is heavily used in national libraries outside the United States [1]. Twenty-four national libraries use LCSH in their national bibliographies. This number does not include the many translations and adaptations of LCSH throughout the world and other familiar subject heading systems based on LCSH. When we turn our attention to LCC, we see that it has become increasing available and accessible. Any of us who have worked with large-scale bibliographic or classification data recognize what an enormous accomplishment it was to convert LCC to the MARC classification. Not only its conversion, but also the way it has been made useable in Classification Plus are accomplishments that the Library of Congress can be proud of. Turning to Dewey, we find that the DDC is sometimes thought of only as a scheme used by public and school libraries, but when we look outside the U.S. we see that DDC is the most widely used classification scheme. The DDC is used in more than 135 countries and has been translated into more than thirty languages [2].

Subject Access on the Web

Next, I would like to turn my attention to subject access on the web. Despite the demonstrated value of our authorized schemes, the application of these systems to web resources is minimal. Quoting Marcia Bates, Lois reminds us how controlled vocabularies provide the consistency, accuracy and control that enables the efficient discovery and retrieval of resources in libraries. Yet despite these benefits, library subject access systems are used only in a very small way on the web.

To investigate the application of two of these systems, I analyzed the application of classification numbers to electronic resources in the CORC database. I looked at DDC and LCC usage in that database eliminating the NetFirst records since they all have Dewey numbers assigned to them. I found approximately 98,000 uses of DDC and 85,000 uses of LC Classification numbers. Although this is a sizable number of records which represents a considerable amount of effort by librarians and other metadata specialists, the use of library categorizations for organizing web resources is essentially invisible on the web. These records and similar ones in library catalogs are considerably less accessible than web sites indexed by Internet search engines and directory services.

When you compare the CORC project to the European subject gateway projects, of which there are many, you find a similar commitment to identifying and describing web resources using standard subject schemes and metadata standards (e.g., Dublin Core). The subject schemes used, include the Ei thesaurus, Agrovoc thesaurus, MeSH, UDC, DDC, etc. It is important to note that, although these gateways are openly accessible on the web, no dominant subject approach has emerged.

Research and Development

Over the past 10 years our research, development and standards efforts have been largely focused on making these schemes easier for humans to use and apply. There are many accomplishments in this regard:

- MARC Format for Classification
- Conversion of LC Classification Schedules
- Explicit coding of form subdivisions
- Introduction of subdivision authority records
- Web version of DDC introduced

Through these efforts, we have made LCSH, LCC, and DDC easier to use, but as a whole, the web community has not embraced our schemes. The web community does not understand library subject systems, has little knowledge of them and what is known, is often based on misinformation. Library schemes are perceived to be outmoded, out-of-date, and only useful for print and older materials

To overcome these biases, we will need to reengineer and re-conceive our schemes for new uses, including

- [Automatic] indexing/categorization
- Surfing vs. searching
- Navigation
- Providing alternate views
- Presenting search results

Lois discusses several adaptations and new uses for LCSH, including a faceted LCSH that may be better suited to the requirements of electronic resources. The DDC is also being used for non-traditional applications in the networked environment. An XML version of Dewey is being used in a pilot project to provide high-level browsing across several subject gateways [3]. To encourage such experimentation and exploration, multiple representations of these schemes will be needed, such as MARC, XML and RDF. An example of a Dewey record in XML is shown below:

```
<?xml version="1.0"?>
<!-- Copyright (C) 2001 OCLC Online Computer Library Center, Inc. -->
<!-- All rights reserved. -->
<rec>
<en><a><ddc>006.31</ddc></a></en>
<eh><a>*Machine learning</a></eh>
<nin><a>Including</a><b>genetic algorithms</b></nin>
<nse><a>For</a><b>machine learning in knowledge-based systems</b><c>,
see</c><d><ddc>006.331</ddc></d><t>.</t></nse>
<nfx><f>*</f><a></a><b>Use notation <ddc>T1--019</ddc> from Table 1 as modified at
<ddc>004.019</ddc></b><t>.</t></nfx> <ieh><a>Genetic algorithms</a><b>computer
science</b><b>artificial intelligence</b></ieh>
<ieh><a>Machine learning</a></ieh>
<SM><f0>sh 94004662</f0><a>Computational learning theory</a></SM>
<FM><f0>sh 94004662</f0><a>Computational learning theory</a><b>--Congresses</b></FM>
<SM><f0>sh 91000149</f0><a>Computer algorithms</a></SM>
<FM><f0>sh 91000149</f0><a>Computer algorithms</a><b>--Congresses</b></FM>
<SM><f0>sh 92002377</f0><a>Genetic algorithms</a></SM>
<EM><f0>sh 96010308</f0><a>Genetic programming (Computer science)</a></EM>
<SM><f0>sh 96010308</f0><a>Genetic programming (Computer science)</a></SM>
<FM><f0>sh 96010308</f0><a>Genetic programming (Computer science)</a></FM>
<SM><f0>sh 85079324</f0><a>Machine learning</a></SM>
<FM><f0>sh 85079324</f0><a>Machine learning</a></FM>
<FM><f0>sh 85079324</f0><a>Machine learning</a><b>--Congresses</b></FM>
<SM><f0>sh 90001937</f0><a>Neural networks (Computer science)</a></SM>
<SM><f0>sh 92000704</f0><a>Reinforcement learning (Machine learning)</a></SM>
</rec>
```

I will conclude with an example that shows how the Dewey classification can be employed in another nontraditional way-to categorize search results. This example was inspired by a talk given by Susan Dumais, a researcher from Microsoft [4]. She and her colleagues evaluated two basic interfaces for structuring search results, a category interface and a list interface. The interfaces were developed to investigate the cognitive processes that lead to effective analysis of results. In the category interface, search results are organized into hierarchical categories and in the list interface, search results are presented as a ranked list. Automatic text categorization was used to categorize the web pages into a broad set of categories based on the categories used on the LookSmart site [5]. At the ASIS&T SIG/CR Classification Research Workshop, Dumais reported that users were not hampered by misclassified items or when results were presented in multiple categories. The sites that could not be categorized were presented in a NotCategorized group.

Through user studies, the researchers found that users preferred the category interface and performed 50% faster at finding relevant information. These results underscore the statements of other speakers at this meeting who called for a greater tolerance for inconsistency or dissonance in our own processes.

To explore whether a similar approach might work in the library environment, I searched in the CORC catalog for the term "cookies." As you can imagine, such a term has multiple meanings. I choose CORC because the resource catalog contains a mixture of traditionally cataloged materials and materials under looser bibliographic control, i.e., DDC numbers assigned using automatic classification [6]. For every record that had a Dewey number, I mapped the number up to its three digit Dewey number. What you see in the example below, is a portion of the search results presented using Dewey categories at the third level of hierarchy.

Data processing Computer science

1. FTP Site of NeoSoft. The FTP site at ftp://ftp.neosoft.com contains the NeoSoft archives. This FTP site is run by NeoSoft, Houston, Texas, in the USA, in a time zone -6 hours from GMT. To access this site over the Web, use URL ftp://ftp.neosoft.com/. The FTP server runs on the UNIX operating system. It also goes by the name of uuneo.neosoft.com.
2. Privacy.net. Features Privacy.net, which provides information about privacy and the Internet, compiled by Consumer.net, a consumer information organization. Discusses cookies, information gathering, encryption, and more.

Computer programming, programs, data

3. Misc.kids Frequently Asked Questions (FAQs): Allergies and Asthma and Recipes. Features recipes for people with allergies and asthma. Notes that the recipes are part of the FAQ section on allergies and asthma of the misc.kids newsgroup. Explains that the information in the FAQ is not intended to replace medical advice. Lists wheat and gluten free recipes for bread, muffins, pancakes, cakes, cookies, and desserts. Provides milk and egg free recipes for cakes, cookies, and desserts. Links to the FAQ section, allergy and asthma resources, and book reviews.
4. Cookies
5. Web Developer's Library (WDVL): Webmaster's Lexicon. Presents a glossary of terms useful to webmasters as part of the Web Developer's Virtual Library (WDVL). Allows users to select individual terms or letters of the alphabet to search for definitions. Defines ActiveX, background, cookies, database, FAQ, graphic, and many other terms. Links to a Web authoring guide, tutorials, a FAQ section, and other Web design-related sites.
6. Programming in JavaScript. Volume two

Food and drink

7. CookieRecipe.com. Presents recipes for all types of cookies. Includes recipes for bar cookies, Christmas cookies, drop cookies, filled cookies, International cookies, molded cookies, no bake cookies, refrigerator cookies, rolled cookies, sugar free cookies, eggless cookies, and gluten-free cookies. Contains a site search engine. Offers conversion tables for common ingredients, as well

as tips and hints. Allows the user to participate in a recipe exchange and submit requests for recipes. Provides a weekly listing of the ten most popular recipes. Links to BreadRecipe.com, PieRecipe.com, and CakeRecipe.com.

8. Egg-stra Delicious Recipes Just for Easter. Presents a collection of Easter recipes. Includes recipes for candy, cakes, cookies, rolls, and cupcakes. Links to other recipe and Easter related Web sites.
9. Cookies and Bars. Features an index of recipes for various cookies and bars. Lists the recipes in alphabetical order. Includes cookies and bars such as snickerdoodles, baklava, biscotti, brownies, gingerbread cookies, lemon bar cookies, and others.
10. Misc.kids Frequently Asked Questions (FAQs): Allergies and Asthma and Recipes. Features recipes for people with allergies and asthma. Notes that the recipes are part of the FAQ section on allergies and asthma of the misc.kids newsgroup. Explains that the information in the FAQ is not intended to replace medical advice. Lists wheat and gluten free recipes for bread, muffins, pancakes, cakes, cookies, and desserts. Provides milk and egg free recipes for cakes, cookies, and desserts. Links to the FAQ section, allergy and asthma resources, and book reviews.
11. M&M's Chocolate Mini Baking Bits. Presents information on M&M's Chocolate Mini Baking Bits from Mars, Inc. Includes recipes for baking with the bits and several hints for successful baking on topics such as choosing butter or margarine, measuring ingredients, preheating the oven, selecting baking sheets, preparing baking sheets, sizing and shaping cookies, storing baked goods, and freezing baked goods. Provides access to a tour of the manufacturing process of the bits. Contains a FAQ section.

If you were to translate the labels into Dewey numbers, the first one is 004, the second is 005 and third one on the page is 641. The same set of results (54 records) that would normally have appeared in a ranked list is shown here broken down by Dewey categories. In this example, resources about Internet cookies appear in the first two categories and resources about the cookies that we like to eat are in the food and drink category. One resource made it into both types of categories. That was one of those dissonant records. In spite of that, the results are promising and suggest that new applications of traditional schemes are possible and that additional experimentation is needed.

References

1. Heiner-Freiling, Magda (2000). Survey on Subject Heading Languages Used in National Libraries and Bibliographies. Cataloging and Classification Quarterly. 29 (1 / 2): 189-198.
2. About Dewey and OCLC Forest Press. Available at http://www.oclc.org/dewey/about/index.htm
3. Renardus. Available at http://www.renardus.org/index.html.
4. S. T. Dumais, E. Cutrell and H. Chen. Classified displays of web search results. Invited presentation at ASIS&T SIG/CR Classification Research Workshop, Nov 12, 2000. Available at http://uma.info-science.uiowa.edu/sigcr/papers/sigcr00dumais.doc
5. LookSmart. Available at http://www.looksmart.com/
6. OCLC CORC / About CORC. Available http://www.oclc.org/oclc/corc/about/corc_over.htm

---

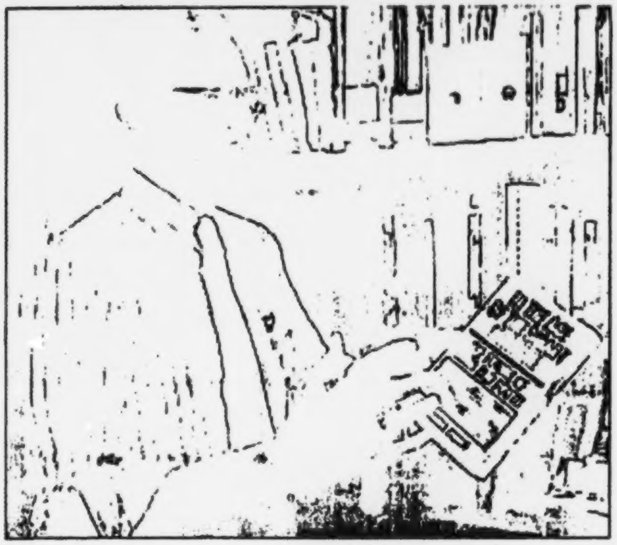
Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

- Conference Home Page
- What's new
- Greetings from the Director for Cataloging
- Topical discussion groups
- NAS study and 2 articles from the LC staff *Gazette*
- Conference program
- Speakers, commentators, and papers
- Conference sponsors
- Conference discussion list
- Logistical information for conference participants
- Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate



William E. Moen, Ph.D.

Assistant Professor
School of Library and
Information Sciences
University of North Texas
P.O. Box 311068
Denton, TX 76203-1068

Resource Discovery Using Z39.50: Promise and Reality

About the presenter:

Dr. William Moen teaches courses in the School of Library and Information Sciences, University of North Texas, on the information of organization, metadata and networked information organization and retrieval, and Z39.50 His research program includes the organization of networked resources, interoperability testing; distributed searching and networked information retrieval, metadata and networked information evaluation, user studies related to networked information seeking behavior, and the development and implementation of technical standards. Since the early 1990s, he has been involved in a number of Z39.50 initiatives including the development of the Government Information Locator Service (GILS) and its associated Z39.50 profile, the coordination of a Z39.50 profile for search and retrieval of cultural heritage information for the Consortium for the Computer Interchange of Museum Information (CIMI), and a comprehensive evaluation of the U.S. Federal implementation of GILS. Most recently, Moen facilitated the development of a statewide Z39.50 profile for Texas (the Z Texas Profile) and participated on the Bath Profile,

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

and international Z39.50 specification for Library Applications and Resource Discovery. He serves as editor of the CIMI, Z Texas, and Bath Profiles. Recent writings include: *Assuring Interoperability in the Networked Environment: Standards, Evaluation, Testbeds, and Politics* (forthcoming) and articles on Z39.50 and related issues published in the *Journal of Academic Librarianship*, *Journal of the American Society for Information Science*, *Texas Library Journal*, and *Communications of the ACM*. Dr. Moen received his Ph.D. from Syracuse University where he wrote his dissertation on the development of the Z39.50 standard.

Full text of paper is available

Summary: The ANSI/NISO Z39.50 protocol for information retrieval is considered by some as an important strategic tool for providing integrated access to distributed networked resources in the while others consider it to be an outdated "technology" that should be abandoned. An understanding of its historical development is critical to evaluate the current perceptions and misperceptions of the roles it is assuming in the networked environment. This paper briefly reviews the 20+ history of Z39.50 development, the complexity of information retrieval problems it addresses, and how the goals for its use has changed over time. In part, the paper shows how this standard was intended to solve problems within a limited community (i.e., libraries) but has now become deployed in other communities to solve the challenges of networked information retrieval. The standard can be viewed as a class of evolutionary standards, and it has evolved to incorporate advances in technologies and technical approaches (e.g., the use of the Internet, integration into the Web environment, and use of new technologies such as the Extensible Markup Language).

The context of Z39.50's goals provides a way to investigate the meaning of resource discovery. Like many terms in the networked environment, resource discovery has many meanings, and the paper attempts to identify the type of resource discovery enabled by Z39.50. Networked resource discovery implies the use of one system to discover resources on one or more separate systems, and such interworking of two systems highlights the key issue of interoperability.

One constant goal of Z39.50 developers was to enable interoperability between diverse systems and diverse resources. The paper describes how Z39.50 enables this interoperability yet details reasons why implementations of the standard have been deficient in achieving this important goal. Recent initiatives have resulted in important national and internationals specifications for using Z39.50 (i.e., profiles) to address underlying interoperability problems, and profiles appear to offer a realistic solution path for seemingly intractable problems in interoperability. The paper describes these profiles and the likely impact they will have on the use of Z39.50 both within libraries and within other communities such as museums. In addition, the paper suggests a framework for analyzing the complexity of interoperability and identifies an approach being developed at the University of North Texas for establishing a rigorous interoperability

testbed.

The past several years has seen a new uptake of Z39.50, both within the library community for creating virtual union catalogs as well as in other communities to solve networked information retrieval problems and provide services to customers. The paper highlights several of these developments to indicate potential roles for Z39.50 in the networked environment. The paper concludes with an overall assessment of Z39.50 strengths as well as the opportunities and challenges the standard faces in serving as a strategic information retrieval tool for libraries and other communities in the networked environment.

Z39.50 continues to evolve as a comprehensive international standard designed to improve the information retrieval of networked resources in a distributed environment, with examples of numerous "profiles" that have been developed over the last several years. This presentation addresses the perception that the standard lacks the broad Internet community support and the contention that it is too flexible and too large and complex for widespread commercial application. It identifies outstanding problems and looks at how well positioned the standard is to offer a future solution to increasing retrieval problems of networked resources on the Web.

Library of Congress
May 9, 2000
Comments: lcweb@loc.gov

Resource Discovery Using Z39.50: Promise and Reality

William E. Moen
Assistant Professor
School of Library and Information Sciences
University of North Texas
Denton, Texas 76203

Final version

The ANSI/NISO Z39.50 protocol for information retrieval addresses the complex challenges of intersystem communication. Original uses envisioned for the protocol look very little like current implementations and uses. In the 1980s, users on one library catalog system would search and retrieve bibliographic records on a remote system. By the late 1990s, there was a need for discovering networked resources and integrating access to them. Yet, the Z39.50 protocol has addressed both these scenarios. This paper provides a portrayal of Z39.50 that explains its flexibility in response to a variety of information retrieval requirements in the networked environment.

What Is Z39.50 Really?

At its most basic, Z39.50 is a communications protocol that enables two systems to exchange messages for the purpose information retrieval. However, one can define and characterize Z39.50 in a number of ways. To begin to understand the use of Z39.50 today, it is worth a brief look back over its 20+ year history [1]. Z39.50 was a realization of 1970s visions for connecting computer systems of large bibliographic utilities and research libraries via telecommunications for purposes of resource sharing, specifically, for sharing MARC bibliographic and authority records. Library leaders such as Henriette Avram saw the potential for resource sharing through the convergence of telecommunications and computers, thus moving towards a regime of national bibliographic control. The National Information Standards Organization (NISO) [2] established Subcommittee D in 1979 to develop a "computer–to–computer protocol for electronic communication of digital information over a network" to support "information transfer at the application level" and would depend on other standards for underlying protocol layers [3]. The Subcommittee focused its initial effort on a protocol for information retrieval.

An Evolving Context for the Protocol

Technical standards can be viewed as solutions to problems. In the case of Z39.50, one can ask what problem was being addressed by the information retrieval protocol. Libraries were the context for the problem. The problem was how to get diverse library automation systems and their underlying information retrieval systems to communicate and thus enable users of one system to search another library's catalog and retrieve MARC records. In its origins, the protocol was intended to solve library problems.

Through the 1980s as the standards committee continued its work, the centrality of the library problem for intersystem communication remained paramount, but new voices became stronger in response to the emerging information retrieval protocol. These voices (e.g., from the abstracting and indexing services) called for a more generalized information retrieval protocol, not one focused only on the intersystem communication between libraries' bibliographic record systems.

With the approval of Z39.50 Version 3 in 1995, the range of implementors of and applications for Z39.50 broadened to include communities with requirements for information retrieval among diverse and distributed resources. Government information, geospatial information, and museum information were three application areas adopting and adapting Z39.50 to the needs of their communities. No longer was the library catalog the central application area for Z39.50.

So, what is Z39.50 really? It is a computer-to-computer protocol that enables intersystem communication for the purpose of searching and retrieving information (where the information can be in the form of MARC records, data from geospatial datasets, museum object records, etc.). But that does not explain why a standard that developed in the context of library problems is now used in a variety of other communities and their applications. For that, we need to look at what the standard offers.

Models, Semantics, and Bits on the Wire

Anyone picking up the Z39.50 standard with the goal of learning what it is, what it does, and how it does it is usually disappointed. Instead of clear descriptions of Z39.50's capabilities and practical uses, the reader is confronted by complex and abstract technical descriptions of facilities, services, application protocol data units, parameters, option bits, and ASN.1 structures. Without initiation into this technical language, the document remains opaque. Yet that technical language does more than confound the average reader. It expresses three important components that are central to what Z39.50 is:

- Abstract models of information retrieval activities (e.g., search, retrieval, etc.)
- A language consisting of syntax and semantics for information retrieval that enables communication between systems
- A prescription for encoding search queries and retrieval results for transmission over a network infrastructure.

Focusing on these components allows us to see the strengths and limitations of Z39.50 for networked information retrieval.

A major contribution of the standard is an abstract model of information retrieval [4]. As an abstract model, it is not tied to any specific implementation, database design, or search engine. Wake states that the "complexity of the Z39.50 information retrieval model should be seen as richness that enables this model to describe many retrieval systems" [5]. The components of the model include (see Figure 1):

- Query: the search submitted by the user (for details about the query, see below on semantics) from a client
- Database: the physical or logical repository of records
- Database record: a local data structure within a database
- Result set: a list created by the server of pointers to database records that meet the criteria of the query
- Retrieval record: the data from the local database record formatted for interchange in a syntax understood by both systems.

This model allowed Z39.50 protocol developers to conceptually separate the user interface (for formulating searches and displaying results) from the information server (with its database management system, search engine and algorithms, local record structure, etc.). Z39.50 protocol machinery in the form of Z39.50 clients and servers mediates between two systems as represented in Figure 2. But for this model to be effective in intersystem communication, protocol developers needed to agree on a language that Z39.50 clients and servers would *speak* to carry out information retrieval transactions.

Figure 1
Abstract Model of Information Retrieval

Abstract Model of Information Retrieval

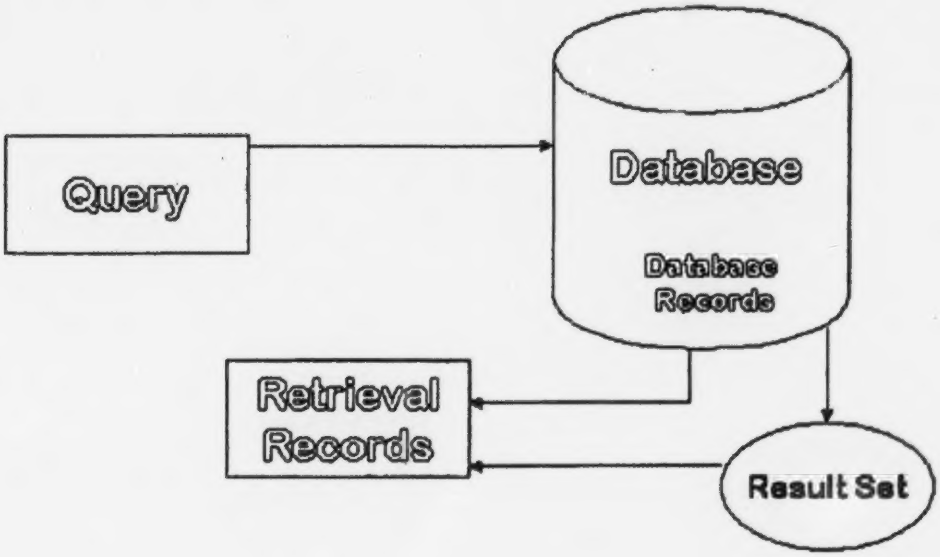
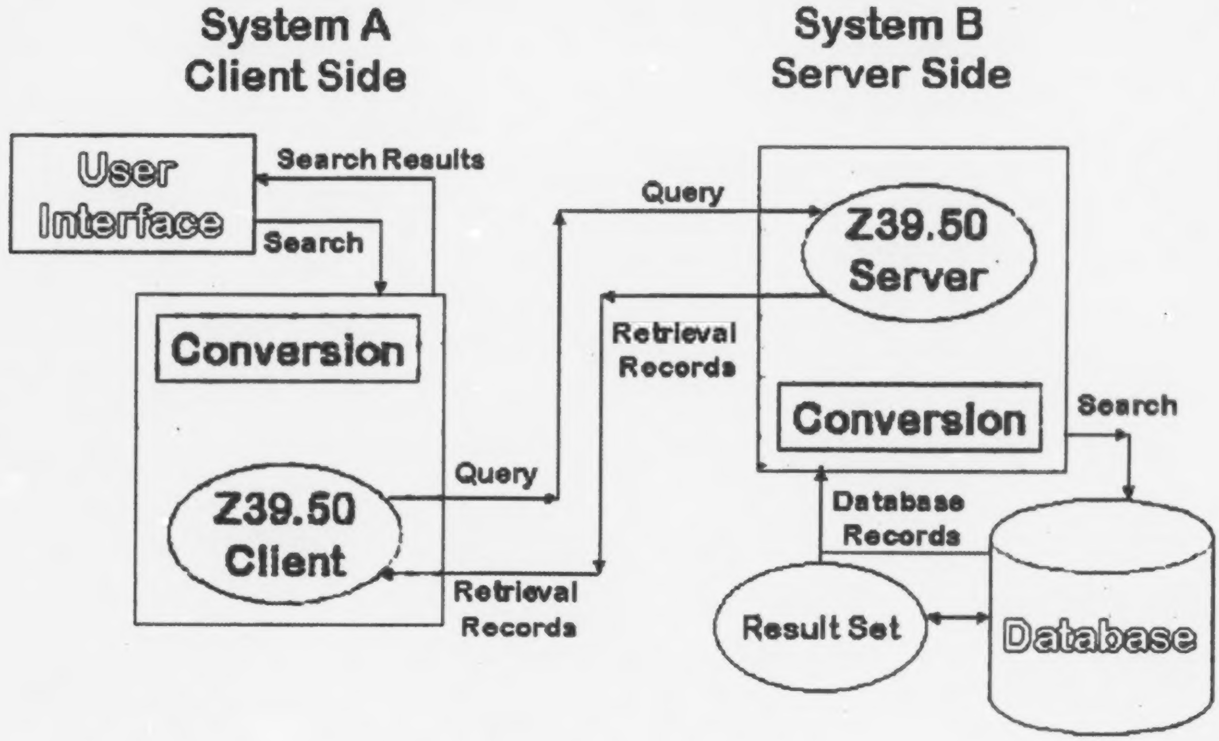

Query
Database
Database Records
Retrieval Records
Result Set

Figure 2
Z39.50 Model of Information Retrieval

Z39.50 Model of Information Retrieval


System A
Client Side
System B
Server Side
User Interface
Search Results
Search
Conversion
Z39.50 Client
Query
Retrieval Records
Query
Z39.50 Server
Retrieval Records
Conversion
Search
Database Records
Result Set
Database

Semantics for Searching and Retrieval

How does a user instruct one system to ask a remote system to do a search for books by Mark Twain? How does the remote system know that the query it receives is requesting a search for books by Mark Twain and not books **about** Mark Twain. What about a title exact match search? What does a title search mean anyway? These questions point to the second major contribution of Z39.50 developers: a semantic model for expressing searches and requesting records that match the criteria of the searches, and the semantics for interchanging the retrieval records.

Each online catalog with its underlying information retrieval system provides users with various search and retrieval options. Typically, search and retrieval options differ between vendors' products. Achieving communication between these disparate systems, each with their own search and retrieval capabilities, was the challenge faced by Z39.50 developers. Getting two systems to exchange protocol messages is one technical challenge, but getting them to "understand" what the messages mean is the arena of semantic interoperability [6].

Building on the abstract model in Figure 1, the developers first worked on standard semantics for expressing queries. More recently, Z39.50 developers focused on semantics and structures for retrieval in a networked information world no longer populated with MARC records. We focus here on semantics for searching to illustrate how Z39.50 addresses semantic interoperability.

In an online catalog environment, users interact with the information retrieval system through an interface where they first formulate their search into a query understood by the machine. A query typically has a search term that is characterized by qualifiers. For example, a search for books by Mark Twain is formulated into a query where the search term is "Mark Twain" (or possibly "Twain, Mark"), and this term is characterized as an "author" term (i.e., search the access point "author"). The qualifiers for the search term tell the information retrieval system how to execute the search: do a search for all records in your database where author is equal to "Mark Twain." We can more precisely characterize the search term and how we want the query executed by additionally describing:

- the structure of the search term (is it a word, a phrase, a date, etc.)
- whether truncation should be performed and if so why kind (no truncation, right truncation, left truncation, etc.)
- whether the search term match the entire field value or only part of the field.

To generalize based on this understanding of what queries are and do, Z39.50 provides attributes sets for expressing searches. Attribute sets define the types of qualifiers available for a search term, and define specific values for those attribute types. For example, the Bib-1 Attribute Set is widely used to express Z39.50 queries against library catalogs. It defines six Attribute Types, each designated by a name and integer: Use(1), Relation(2), Position(3), Structure(4), Truncation(5), and Completeness(6). Each attribute type can take on values (also designated by name and integer value). For example, a Use attribute characterizes the access point that should be searched. One Use attribute value is "Title" or "4" to designate a title access point. Attribute types and values are expressed as integer pairs; the pair (1,4) tells the server to execute a title search. The combination of attribute types and values provides a way to express the semantic intention of the search and prescribe the behavior expected when the server executes the query. For example, we can express a ***keyword author search*** for Twain as ***(1,1003) (2,3) (3,3) (4,2) (5,100) (6,1) Twain***, where:

- Use Attribute (1) = author (1003)
- Relation Attribute (2) = equal (3)
- Position Attribute (3) = any position in field (3)
- Structure Attribute (4) = word (2)
- Truncation Attribute (5) = do not truncate (100)
- Completeness Attribute (6) = incomplete subfield (1).

I've illustrated in some detail how Z39.50 addresses semantic interoperability for searching by providing a standardized language (syntax and semantics) for expressing queries. For meaningful communication to occur, the communicating Z39.50 client and server must "know" or recognize values from a common attribute set (e.g., Bib-1). Only then will they be able to meaningful exchange and process a query. For example, the client will be able to convert a search expressed in the structure of its local information retrieval (IR) system into standard Z39.50 vocabulary; and the server will be able to receive and understand the Z39.50 query and convert it into its local IR system search logic for execution. Figure 2 indicates the conversion points for mapping into and out of the Z39.50 protocol language on the client and server.

The expressiveness offered in Z39.50 for queries grew out of the context for the protocol, namely, searching large online catalogs and bibliographic databases accessible by robust information retrieval systems. These databases held well-structured bibliographic records created according to national and international standards and guidelines. The information retrieval systems provided any number of access points to the records including author, title, and subject, and allowed the end-user to qualify and refine searches to improve retrieval results. The model for searching was not simple keyword access. Z39.50 functionality mirrors the search and retrieval functionality of those online library catalog systems. One power of Z39.50 is being able to communicate precision-oriented (as well as recall-oriented) searches against well-structured information in the form of bibliographic records or other forms of structured metadata. What are the implications of this for resource discovery?

Resource Discovery

We know that resource discovery must be a good thing since lots of people want to do it and many claim they have tools to do it. Like the term metadata, resource discovery has many connotations. To evaluate the use of Z39.50 for resource discovery, it is helpful to have a working definition of the concept. Lynch suggests that the resource discovery is used to describe a complex collection of activities, from "simply locating a well-specified digital object on the network all the way through lengthy iterative research activities....Discovery often involves the searching of various types of directories, catalogs, or other descriptive databases....Most often, the discovery process operates on surrogates (such as descriptions) of actual networked information resources" [7]. Key elements of resource discovery appear to be finding, identifying, and accessing information, and the use of representations or surrogates in the discovery process.

Lynch characterizes networked information resources as "digital objects, collections of digital objects, or information services on the network" [7]. One can use the Internet to discover all kinds of resources, such as people, organizations and institutions, products, services, texts, images, sounds, and so on. Each of these resources are represented digitally in some fashion. People could be represented by the occurrence of their name on a document, in an email message, or on a website. Organizations and institutions might be represented by a company website. How these objects are represented will likely determine the utility of Z39.50 for discovering them.

From the perspective of the Z39.50 abstract information retrieval model, there is a database that contains records, where a records is a surrogate for some thing (e.g., a digital object). With Z39.50, a Z39.50 client knows of the existence of a Z39.50 server (e.g., network address, port number, etc.) and possibly names of one or more databases made accessible via the server. This means that to get started with resource discovery using Z39.50, a client must know at least one server. But that is really no different than needing to know the URL for AltaVista or Google to get started doing resource discovery using Web search engines.

Apples and Oranges, Search Engines and Z39.50

One can hardly discuss networked information discovery and Z39.50 without a brief discussion of web search engines. Although it is critical in evaluating Z39.50 role in resource discovery to clarify the differences between Z39.50 and web search engines, the scope of this paper does not allow an extended treatment. Z39.50 is an intersystem communications protocol for information retrieval. It is not a search engine. A Z39.50 client can send searches to one or more database on remote systems at the same time (from the perspective of the user). It allows the user to see these different databases as if they were one logical resource. The client connects with each separate server, searching the current contents of the database, and getting results directly from the source databases. Z39.50 simply provides the protocol for these systems to communicate information retrieval messages. One can characterize this approach to networked information retrieval as decentralized or multi-system.

A web search engine is fundamentally a single information retrieval system that has the added function of harvesting resources from the Internet and performing some sort of indexing to make those resource searchable. When users are interested in discovering resources via a web search engine, their web browser presents a search interface for that search engine, and a query is executed against the databases and indexes of that single search engine. One can characterize this approach to networked information retrieval as centralized or single-system.

The stored representations may differ significantly between a Z39.50 accessible database and the web search engine databases. In the latter, the harvested networked information resources are typically represented by words/terms taken from the document and placed in a index. There is no structured representation for the resources. Z39.50 accessible databases typically contain structured representations or surrogates for the resources. These may be in the form of library catalog bibliographic records, museum object records, collection-level records, or other forms of structured metadata.

Granularity and Aggregation: What are Users Trying to Discover?

We noted above that a Z39.50 client must "know" about a Z39.50 server prior to getting started. There are published lists of Z39.50 servers, but the larger challenge is selecting an appropriate server for a particular information need. Subject gateways, such as the Arts and Humanities Data Service [8], assist users by identifying a number of resources (i.e., databases) that are Z39.50 accessible and provide a Web search interface for using Z39.50 to search one or more of the identified resources at the same time. The gateway is a logical aggregation of several discrete networked information resources. This raises the question as to what the resources are that discovery tools are helping users discover? Web search engines work at the level of an HTML file (the addressable unit for retrieval), where the file can be a report, a homepage, a poem. Z39.50 models resources as records (the addressable unit for retrieval in a database), where the record can represent almost anything that can be described.

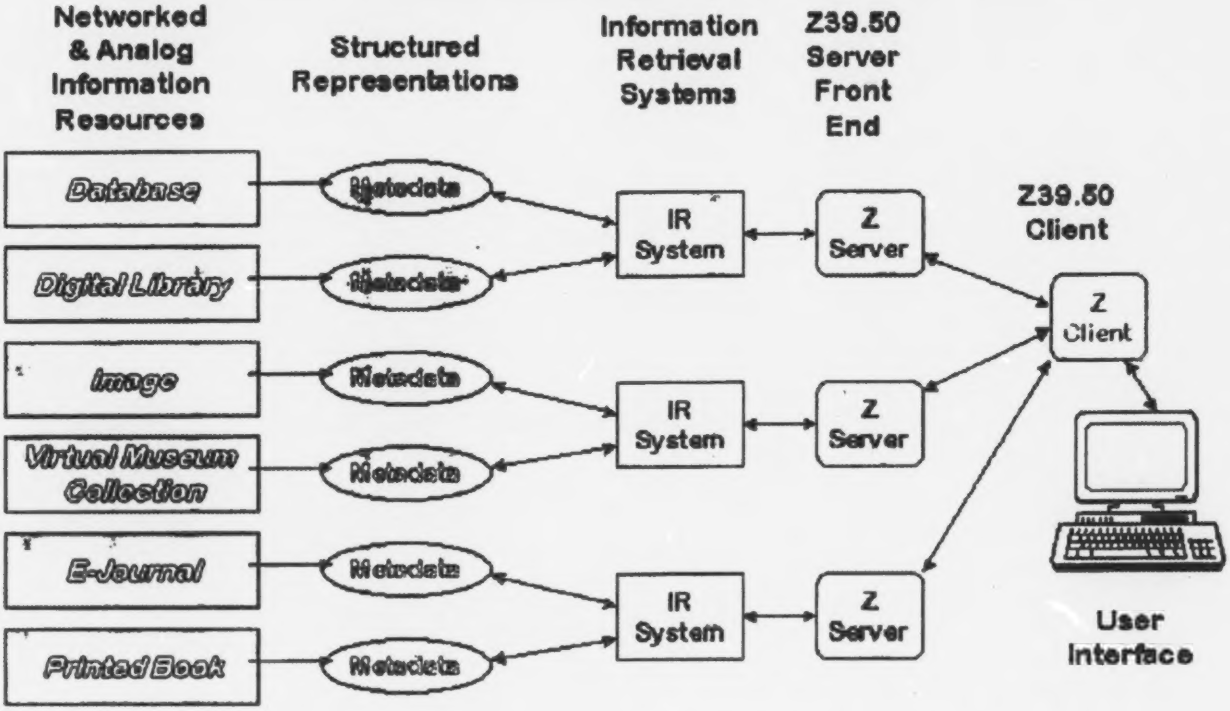
The library cataloger's concept of unit of analysis (or unit of description or unit of retrieval) is useful in this context. This concept helps catalogers identify what exactly they are representing in a single bibliographic record. In the print world, various levels of granularity or aggregation can be represented. For example, a single volume of an monographic set can be described in a bibliographic record; the monographic set also can be described.

In terms of resource discovery, what exactly is the *size* or *scope* of the resource we are trying to discover? Are we looking for a web page? A web site? A text document comprising a number of web pages? A specific graphic image that is part of a web page? A database of records? The unit of analysis for web search engines is an addressable file. The unit of analysis for Z39.50 can be anything, but the record-based model for Z39.50 assumes that a resource is represented by a logical, if not a physical, record. Some examples can illustrate this. In a library catalog, a record might represent an item in a library's collection such as a book, journal, map, etc. But a record might also represent a series, a set of items. In an abstracting and indexing service (A&I) database, a record might represent a journal article. In a museum collection management system, a record might represent a specific art object. We can categorize all of these as metadata records, structured records that describe resources. We can also envision descriptive metadata records created to represent an online database, a repository of electronic texts, a museum and collections housed by that museum. This moves us to a context in which Z39.50 can be viewed as a tool for resource discovery. As long as the resources are represented and made available through some sort of information retrieval system, those resources could be discovered via Z39.50. Figure 3 illustrates a Z39.50 client accessing one or more Z39.50 accessible information retrieval systems that have records representing information resources. Z39.50 discovers those resource descriptions. Whether or not the described resources are accessible via Z39.50 or any network tool is another issue.

To accomplish Z39.50 resource discovery, the system represented by the User Interface in Figure 3 must be interoperable with one or more remote information retrieval systems and the databases served by those information retrieval systems so meaningful communication occurs. The challenge is, can a user formulate a search using a Z39.50 client to search one or more remote systems and get meaningful results? This is the fundamental challenge of interoperability.

Figure 3
Z39.50 Model of Resource Discovery

Z39.50 Model of Resource Discovery


Networked & Analog Information Resources
Structured Representations
Information Retrieval Systems
Z39.50 Server Front End
Z39.50 Client
Database
Digital Library
Image
Virtual Museum Collection
E-Journal
Printed Book
Metadata
Metadata
Metadata
Metadata
Metadata
Metadata
IR System
IR System
IR System
Z Server
Z Server
Z Server
Z Client
User Interface

Interoperability

Interoperability is a key issue for resource discovery and more generally networked information retrieval [9]. Interoperability is a concept that addresses the extent to which different types of computers, networks, operating systems, and applications work together effectively to exchange information in a useful and meaningful manner. The networked environment is heterogeneous; it hosts many different technologies, various data, multiple applications, and other networked life–forms. A functional goal in this environment is to hide this heterogeneity from users so they may effectively do business, search for information, communicate, and perform other tasks. There is little doubt interoperability is a key issue in the networked environment [6, 10, 11, 12]. Interoperability or its absence can affect information access. Technical interoperability can raise important policy and organizational issues [13].

BEST COPY AVAILABLE

As a working definition of interoperability for this paper is: *the ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner* [14]. Based on experiences with Z39.50 implementations, several levels and types of interoperability can be articulated including:

- **Low–level protocol (syntactic):** do two implementations interchange protocol messages according to the standard?
- **High–level protocol (functional):** do two implementations support the same Z39.50 services as defined in the standard?
- **Semantic level:** do two implementations preserve and act on meaning of information retrieval tasks?

Z39.50 implementation experience gained over the past decade has solved most of the low-level protocol interoperability problems. The high-level protocol interoperability problems are resolved for the most part when a Z39.50 client and Z39.50 server support the same services (e.g., sort, scan). The arena of semantic interoperability is where Z39.50 developers and implementors face the most complex set of challenges.

Semantics for Searching Revisited

We discussed above how Z39.50 provides a language for expressing queries, and this language with its attendant syntax and semantics, enables two systems to *understand* each others requests and responses. In practice this *understanding* has not always been achieved. The lack of semantic interoperability has caused users to lose confidence in Z39.50 interfaces to information retrieval systems (whether their native systems or remote systems). What affects semantic interoperability? The two major factors affecting interoperability are differences in Z39.50 implementations and differences in indexing decisions in the information retrieval systems. The results of these differences show up in retrieval results. Going back to the analogy of Z39.50 as a language, the meaning (semantics) of the protocol messages needs to be clear if two systems are to share an "understanding" of the message. Z39.50 provides standardized "vocabularies" to express queries using registered sets of attributes (where attributes are used in the Z39.50 query to characterize a search term). The attribute sets provide the "words" in the vocabulary for searching.

Z39.50 implementations, however, do not always support (i.e., understand and act on) the same "words" from the standardized vocabulary for searching. Taking an example from library catalogs, System A wants to search System B for a corporate author and formulates the query using the correct Z39.50 attribute type/value pair to characterize its search term as a corporate author. But System B does not support that particular Z39.50 attribute type/value pair. The semantic intention of the user and his/her search cannot be acted upon. However, the System B does support a name search, and in an attempt to be helpful, processes the corporate author search as a name search; the results, however, may include records that are not relevant to the original corporate author search; semantic loss has occurred. In both these cases, semantic interoperability is reduced or does not exist.

The Semantic level of interoperability is also affected by the local information retrieval system's functionality and indexing policies. Although the standard provides mechanisms for clearly—if not unambiguously—expressing search requests, retrieval requests, and other IR functional requests, the differences in local systems can jeopardize semantic interoperability. In the example above, the two systems are online library catalogs (i.e., bibliographic databases) populated with records derived from standard MARC records. However, System A allows specific MARC fields to be searched for corporate author names while System B, with the same basic set of records, has chosen not to create indexes or is incapable of creating indexes to support the access point of corporate author. Thus System B is incapable of doing a search for corporate author even though the Z39.50 server front end to its system can process and understand the query. There is likely a strong relationship of the search capabilities of the underlying IR system and the Z39.50 attributes it supports in its Z39.50 server software. Further, Z39.50 client and server software cannot add functionality to a local IR system that it doesn't have.

As a community, we are beginning to grasp the impact of local systems' functionality, local indexing decisions and policies, normalization practices, etc., on interoperability. These impacts go beyond issues of Z39.50 conformance but part of the interoperability equation can be addressed by Z39.50 profiles.

Z39.50 Profiles: Solutions to Semantic Interoperability

Profiles can be considered auxiliary standards mechanisms. They define a subset of specifications from one or more standards to improve interoperability. The objective of a profile is to detail a set of specifications from options and choices available in a base standard(s) to address specific technical or functional requirements. Implementors' products conforming to a profile have an improved likelihood of interoperability. Two motivations have initiated Z39.50 profiles:

- to prescribe how Z39.50 should be used in a particular application environment (e.g., government information, cultural heritage museums, etc.)
- to solve interoperability problems with existing Z39.50 implementations within a community or across two or more communities (e.g., the library community).

This section discusses how profiles can address semantic interoperability problems in cross-catalog searching.

Between 1999 and 2000, an international effort produced *The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery* [15, 16]. The Bath Profile itself was informed by several previous profiles, but most importantly by the *Z Texas Profile: A Z39.50 Profile for Library Systems Applications in Texas* [17, 18]. These two profiles focused effort on resolving semantic interoperability problems for cross–catalog information retrieval, and they prescribed the specific Z39.50 services required to support various user tasks (e.g., Init, Search, Present, Scan).

In the case of the Bath Profile, it addresses semantic interoperability for searching by defining a core set of 19 searches; requirements for these cross–catalog searches resulted from discussions among librarians. Defining the searches included naming a search, prescribing IR system behavior to process the query, and prescribing the Z39.50 query vocabulary to unambiguously express each defined search. For example, the Profile defines an ***Author Keyword Search with Right Truncation***. The semantics (i.e., prescribed IR system behavior) for that search is: "Searches for complete word beginning with the specified character string in fields that contain the name of a person or entity responsible for a resource." The specification of the query using Z39.50 Attributes is:

- Use Attribute (1) = author (1003)
- Relation Attribute (2) = equal (3)
- Position Attribute (3) = any position in field (3)
- Structure Attribute (4) = word (2)
- Truncation Attribute (5) = right truncation (1)

- Completeness Attribute (6) = incomplete subfield (1).

This combination of attribute types and attribute values expresses this and only this search. Thus, there should not be any ambiguity of what a server is to do when it receives this query. If the Z39.50 server and its database is unable to understand this query or to process it in the way prescribed, it should fail the search and return a diagnostic to the Z39.50 client.

Even though the profiles address the Z39.50 aspect of semantic interoperability, the semantic level is also affected by the indexing policies and search functionality in the local IR system. To address the variations in indexing in different systems, the approach of the Texas Z39.50 Implementors Group (TZIG) is to recommend a common indexing policy to support the searches specified in the Profile. Recommending indexing policies goes beyond the scope of Z39.50 specifications, but to improve semantic interoperability, we have concluded that common indexes populated with data from a core set of MARC fields and subfields is essential.

The library community is quite homogeneous, especially in terms of its catalogs. But the diversity -- in Z39.50 implementations and local information retrieval systems -- is now reducing the ability of users (whether information professionals or end user patrons) to take advantage of the networked environment to discover and retrieve pertinent resources. The experience with the Bath and Z Texas profiles suggest that a new level of standardization and consistency in Z39.50 implementation, information retrieval functionality, and indexing practices is necessary to achieve meaningful networked information retrieval among library catalogs.

Virtual Union Catalogs and Cross-Domain Searching

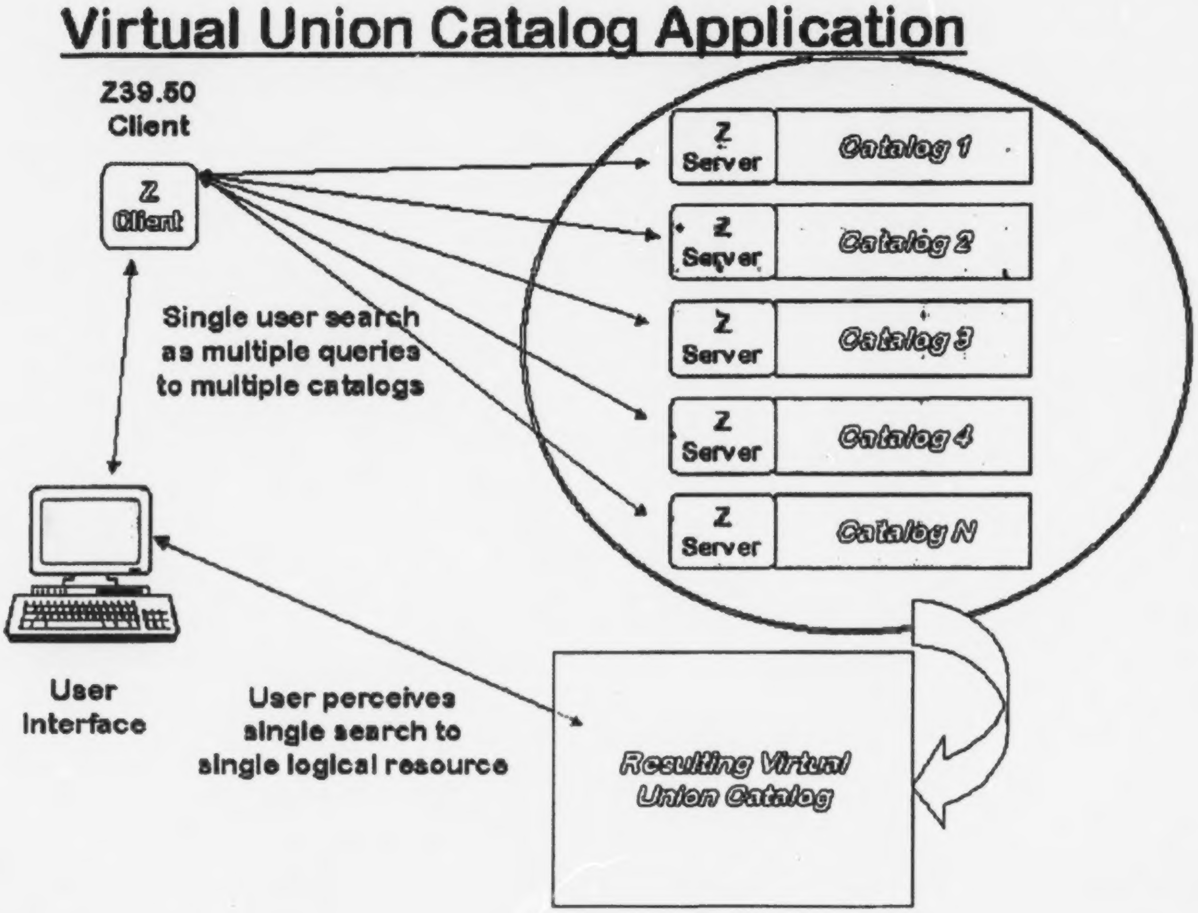
The final sections of this paper present two applications areas in which Z39.50 is being used currently. These fall generally into the arena of resource discovery since these applications involve the identification of an information resource for retrieval and access.

Virtual Union Catalogs

Although the original model of intersystem communication for Z39.50 focused on a Z39.50 client interacting with a single Z39.50 server, implementors in the 1990s began developing clients that allowed a user to interact with more than one Z39.50 server at a time. This gave the user the capability of formulating a single search that would be executed against two or more separate databases. The Z39.50 client established Z39.50 sessions with one or more servers, sent the query to each of those servers, and retrieved results from each server to present to the user. >From the user's perspective, he/she was simultaneously and transparently searching multiple resources at the same time. As a result, the multiple resources being searched at the same time appeared to the user as a single search against one logical resource.

Librarians saw the potential for this in the context of union catalogs [19]. Why not use the distributed searching capabilities of Z39.50 to create virtual union catalogs by virtue of sending the same query to multiple catalogs simultaneously? Would it be possible to abandon the physical union catalog in favor of a virtual union catalog? Figure 4 illustrates how a Z39.50 client connects to multiple, remote catalogs for search and retrieval. A single search from the user is sent to multiple Z39.50-accessible catalogs and results from each catalog are returned. Depending on the client-side capabilities, the results from each of the catalogs could be merged into a single result set with duplicate records removed, etc. From the users' perspective, however, the search goes against a logical resource (i.e., the virtual union catalog) rather than against separate catalogs.

Figure 4
Virtual Union Catalog Application


Virtual Union Catalog Application
Z39.50
Client
Z
Client
Single user search
as multiple queries
to multiple catalogs
Z
Server
Catalog 1
Z
Server
Catalog 2
Z
Server
Catalog 3
Z
Server
Catalog 4
Z
Server
Catalog N
User
Interface
User perceives
single search to
single logical resource
Resulting Virtual
Union Catalog

The use of Z39.50 doesn't mean the end of traditional union catalogs. For example, Clifford Lynch suggests that we should see that the single physical union catalog model "complements the emerging distributed search models by offering substantially different functionality, quality, performance, and management characteristics" [20]. To adequately assess the utility of either model, however, studies are needed to evaluate these differences. Coyle provides one of the first systematic looks in comparing a centralized union catalog (i.e., Melvyl) with a virtual union catalog [21].

Performance issues may become paramount considerations. For example, in a virtual union catalog each search will go to each participating catalog. Smaller public libraries participating in such a catalog may be subject to large numbers of virtual union catalog search that could put an adverse load on local computing resources compared to a large academic library participant with a more robust computing and networking infrastructure. Performance issues have yet to be investigated systematically.

BEST COPY AVAILABLE

And we also have to deal with the ever-present semantic interoperability issues in a virtual union catalog model. Unless each participating catalog's Z39.50 server is configured similarly for support of Z39.50 attribute types and values, and each catalog's indexing policies are similar, users may be less satisfied with the results from a virtual union catalog than from a centralized single union catalog database [19]. These semantic interoperability problems, however, are susceptible to the solutions provided by Z39.50 profiles.

Cross Domain Searching

Library catalogs are not the only resources that are Z39.50 accessible. Efforts in the cultural heritage museum, natural history museum, archives, government information, and geospatial communities to implement Z39.50 solutions for networked information retrieval are making a diverse set of information resources available to Z39.50 clients. It may be that when one thinks of the concept resource discovery, this heterogeneous networked information environment is what captures their imagination. Think of a user with a need for information about the artist Van Gogh. Certainly the user might be interested in discovering books about the artist, but he/she might also be interested in discovering manuscript collections, images, museum collections and exhibits, etc. related to Van Gogh. The user might begin with a search of several library catalogs plus one or more museum systems and an archive or other metadata repository to find relevant information. Librarians and library users desire integrated access to distributed resources where those resources may take different forms (e.g., images, books, sound recordings, etc.). As Hammer noted, "The essential power of Z39.50 is that it allows diverse information resources to look and act the same to the individual user" [22]. Is this, then, really the promise of Z39.50 and resource discovery?

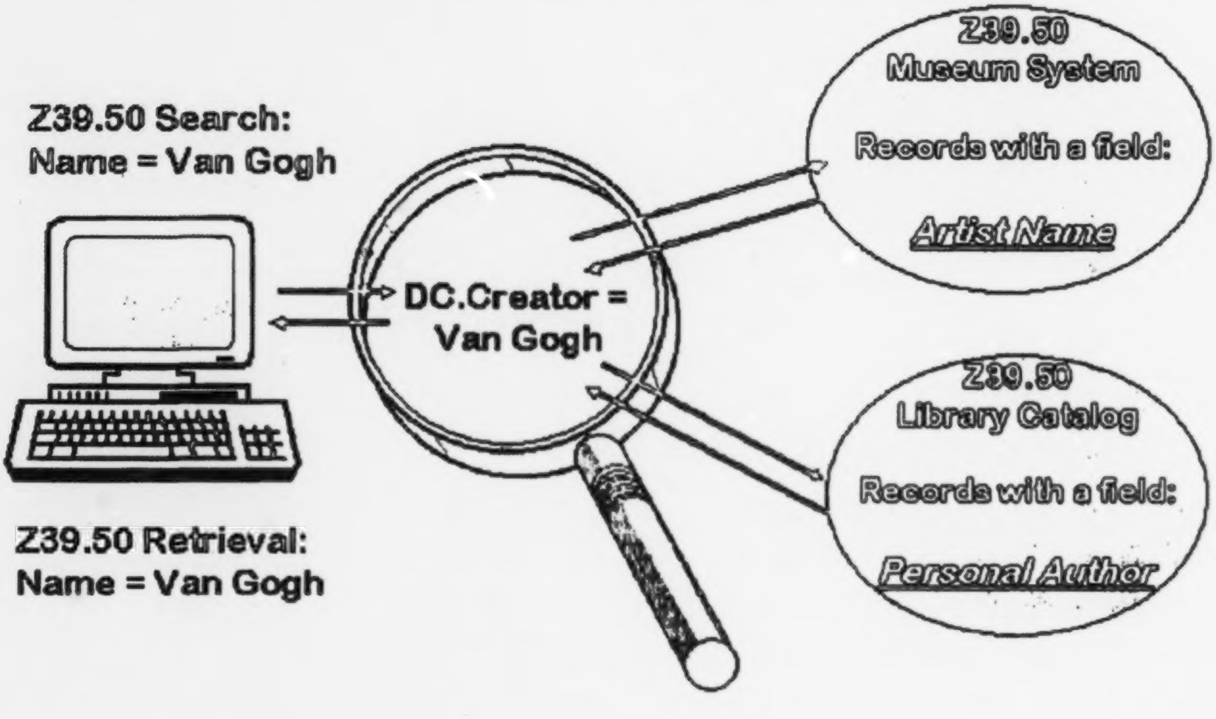
Z39.50 can be used to provide effective cross-domain searching of diverse resources including library catalogs, government information, museum systems, and archives. A library's Z39.50 client configured for cross-domain searching could send out queries to Z39.50 accessible museum and archive systems configured to support cross-domain searching. Similarly, a museum curator could use a museum Z39.50 client configured to support cross-domain searching to search the local museum system, one or more other museum systems, one or more library catalogs, and government resources that are Z39.50 accessible and configured to support cross-domain searching. A project conducted by the Consortium for the Computer Interchange of Museum Information (CIMI) demonstrated how cross-domain searching could be done across library catalogs and museum collections [23].

One mechanism to enhance Z39.50 cross-domain searching is to use the Dublin Core Metadata Elements to provide semantic interoperability for expressing search requests and packaging retrieval results. In the virtual union catalog described above, there is a homogeneity to the bibliographic records in each catalog (e.g., most all records have a concept of author, title, etc.; they can be interchanged as MARC records). When one moves outside a single domain, that homogeneity of semantics and data structures is removed. In a museum's collection management system, the person responsible for the intellectual work of a painting is seldom referred to as an author but more likely as artist. Yet there is a level of semantic equivalence between the concepts *author* and *artist*.

The Dublin Core Metadata Elements address semantic interoperability for resource discovery [24]. The elements themselves can be used as the "words" in the Z39.50 query vocabulary (i.e., as Use Attributes in Z39.50 to be able to characterize search terms). The Dublin Core elements become a lens through which a Z39.50 client sees a wide range of diverse resources. Similarly, an information retrieval system can make its resource visible through the Dublin Core elements. For retrieval purpose, a Z39.50 server can package up a retrieval record using the Dublin Core elements as labels for the units of information or fields of the retrieval record. Figure 5 illustrates how cross-domain searching can be enabled through the use of Dublin Core elements.

Figure 5
Cross Domain IR Application

Cross Domain IR Application


Z39.50 Search:
Name = Van Gogh
DC.Creator =
Van Gogh
Z39.50
Museum System
Records with a field:
Artist Name
Z39.50
Library Catalog
Records with a field:
Personal Author
Z39.50 Retrieval:
Name = Van Gogh

While most of this paper has focused on interoperability issues related to searching, there is an associated set of issues related to retrieval interoperability. In the cross-domain environment, retrieval issues become much more pronounced than in the virtual union catalog. In the latter, retrieval interoperability is achieved through the use of a MARC record syntax for the retrieval record. Most library catalogs can export legitimate MARC records, and these can be passed between the server and client via Z39.50.

Searching across domains, however, offers no such pre-existing standard for a data interchange format. Z39.50 developers addressed this problem in the early 1990s by defining a Generic Record Syntax (GRS) to express arbitrarily structured database records in a standard format for interchange in Z39.50. While this proved to be a viable solution within the Z39.50 community, a more likely solution is the integration of Extensible Markup Language (XML) as a core record syntax for use in Z39.50. Whether GRS or XML, addressing semantic interoperability on the retrieval side is as pressing as the semantic interoperability on the searching side when doing cross-domain searching.

Z39.50's Future in Networked Information Retrieval

The ANSI/NISO Z39.50 protocol for information retrieval is considered by some as an important strategic tool for providing

BEST COPY AVAILABLE

integrated access to distributed networked resources. Others, however, consider it to be an outdated "technology" that should be abandoned. Assessing its utility necessitates a clear statement of the application and functional requirements in which Z39.50 is being considered. Clear functional requirements for an application can then allow us to determine if Z39.50 or some alternative technology is appropriate.

This paper has briefly reviewed the 20+ year history of Z39.50 development, the complexity of information retrieval problems it addresses, and how the goals for its use has changed over time. This standard -- intended to solve problems within a limited community (i.e., libraries) -- now is deployed in a range of other communities to solve the challenges of networked information retrieval. The standard can be viewed as a class of evolutionary standards, and it has evolved to incorporate advances in technologies and technical approaches (e.g., the use of the Internet, integration into the Web environment, and use of new technologies such as the XML).

Where does the perception that Z39.50 represents outdated technology arise? Without some attention to this issue, any discussion of Z39.50's future is clouded. Z39.50's origins in the Open Systems Interconnection (OSI) framework of the 1970s and 1980s have not been forgotten (nor entirely removed from the standard). The power of Z39.50 comes at a cost of complexity. Setting up a web server and full-text indexing search engine is commonplace. How common is it for an operating system to bundle an easy-to-configure Z39.50 server as, for example, Linux does with the Apache web server? Available Z39.50 toolkits may require not only significant C or C++ programming experience but also require familiarity with the less-than-common technical tools such as Abstract Syntax Notional One (ASN.1) and Basic Encoding Rules (BER) to encode the protocol messages for transmission over the wire. A Z39.50 implementor has to address a range of concerns from abstract semantic models to the bits passing over the wire. And, for the most part, there is little off-the-shelf software that can make implementing Z39.50 clients or servers easy to do. Certainly we don't see Z39.50 plug-ins for Netscape and Internet Explorer.

Will Z39.50 be relegated to a backwater of networked information retrieval? It is a standard that addresses important interoperability challenges but does so in a way, perceived as a library way, that may keep it a niche solution rather than as a broader solution to critical problems of networked information retrieval. This paper has argued that major contributions of Z39.50 have been abstract and semantic models for information retrieval. The question is whether and how the Z39.50 community can leverage these contributions while letting go of some of the arcane technical aspects of the protocol that keep it from being widely adopted. At the July 2000 international Z39.50 Implementors Group (ZIG) meeting in Leuven, Belgium, participants agreed to build on the strengths of Z39.50 (the modeling and abstraction) and investigate how other technologies and newer protocols could be used (e.g., SOAP and the emerging XML Protocols).

Z39.50's future in broader networked information retrieval environment is uncertain. The complexity of distributed networked information retrieval is not appreciated until one tries to do it. Information retrieval from a single IR system is not problematic (as is the case with the web search engines). Distributed search across multiple servers with different database systems and different data and semantic structures is problematic. Experience with Z39.50 has identified many aspects of the complexity of distributed search and retrieval. Z39.50 developers and implementors have worked to resolve many interoperability issues, but too often the successes have come slowly and usually not with great fanfare.

The strategy for success being followed by the Bath and Z Texas Profile developers may be considered an incremental strategy. We are trying to rebuild confidence in Z39.50 for a group of users that should not have lost confidence in the first place, namely, librarians. We are not promising that Z39.50 will solve all information retrieval problems. But the profiles offer an opportunity to show how Z39.50 can be used successfully in the original community that developed the standard. Discussing Z39.50's role in resource discovery as compared with web search engines, although attempted in this paper, may be one more tangent from the pragmatic roles for Z39.50:

- as a standard that provides an example of mechanisms for "standardizing shared semantic knowledge" [4]
- as a practical tool in the arsenal of librarians and information professionals in search and retrieval across multiple library catalogs
- as a potential strategic tool for integrating access to selected networked information resources.

Success in these three roles is possible. Demonstrable and effective use of Z39.50 within the library community has not been a given. We can at least start Z39.50's future by making it work for us in the present.

Notes

1. For a history of Z39.50 development, see Moen, William E. (1998). *The development of ANSI/NISO Z39.50: A case study in standards evolution*. Unpublished doctoral dissertation. Syracuse, NY: School of Information Studies, Syracuse University. Available: <http://www.unt.edu/wmoen/dissertation/DissertationIndex.htm>.
2. The National Information Standards Organization (NISO) resulted from an organizational change in 1984. Prior to that point, the name of the standards organization was the American National Standards Committee Z39. This paper will use NISO to reference the current as well as earlier standards organization.
3. American National Standards Committee Z39. (1979, February 2). *Information form: Recommended ANSC Z39 Standard*. Gaithersburg, MD: American National Standards Committee Z39.
4. Lynch, Clifford A. (1997, April). The Z39.50 information retrieval standard: Part I: A strategic view of its past, present and future. *DLib Magazine*. Available: <http://www.dlib.org/dlib/april97/04lynch.html>.
5. Wake, William. (2000, February). *Analysis objective: Z39.50 search system - object model.* Available: <http://users.vnet.net/wwake/analysis/z3950/z3950.shtml>.
6. Semantic interoperability is viewed as a difficult problem to solve. For example, Clifford Lynch and Hector Garcia-Molina stated: "Deep semantic interoperability is a 'grand challenge' research problem; it is extraordinarily difficult, but of transcendent importance, if digital libraries are to live up to their long-term potential." Lynch, Clifford and Garcia-Molina, Hector. (1995). *Interoperability, scaling, and the digital libraries research agenda: A report on the May 18-19, 1995 IITA digital libraries workshop.* (U.S. Government's Information Infrastructure Technology and Applications (IITA) Working Group, Reston, Virginia). Available: <http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/>.
7. Lynch, Clifford A. (1995, October). Networked information resource discovery: An overview of current issues. *IEEE Journal on Selected Areas in Communications, 13*(8): 1505-1522.
8. Arts and Humanities Data Service. (2000). *The AHDS gateway.* Available: <http://ahds.ac.uk:8080/ahds_live/>.
9. For a longer exposition on interoperability in the context of Z39.50, see Moen, William E. (forthcoming). "Assuring interoperability in the networked environment: Standards, evaluation, testbeds, and politics. In McClure, Charles R. and Bertot, John Carlo, Eds. Evaluating networked information services: Techniques, policy, and issues. Silver Spring, MD: American Society for Information Science, Information Today, Inc.
10. Lynch, Clifford. (1993, March) Interoperability: The standards challenge for the 1990s. *Wilson Library Bulletin, 67*(7): 38-42.
11. Payette, Sandra, Blanchi, Christophe, Lagoze, Lagoze, Overly, Edward A. (1999, May). Interoperability for digital objects and repositories. *D-Lib Magazine, 5*(5). Available: <http://www.dlib.org/dlib/may99/payette/05payette.html>.
12. Miller, Paul. (2000, June 21). Interoperability: What is it and why should I want it? *Ariadne* 24. Available: <http://www.ariadne.ac.uk/issue24/interoperability/intro.html>.
13. Moen, William E. (2000). Interoperability for information access: Technical standards and policy considerations. *The Journal of Academic Librarianship, 26*(2): 129-132.
14. Abbas, June, Monika Antonelli, Mark Gilman, Pamiela Hight, Valli Hoski, Jodi Kearns, Teresa Lepchenske, Martha Peet, Mike Pullin, and Amy Stults. (1999). *An Overview of Z39.50, supplemented by a case study of implementing the Zebra server under the Linux operating system.* Denton, TX: School of Library and Information Sciences, University of North Texas.
    Available: <http://www.unt.edu/wmoen/Z3950/GIZMO/contents.htm>.
15. Bath Profile Group. (2000, June). *The bath profile: An international Z39.50 specification for library applications and resource discovery, Release 1.1. An internationally registered profile.* Available: <http://www.ukoln.ac.uk/interop-focus/bath/current/>.
16. For background on the Bath Profile see Lunau, Carrol. (2000, March). *The Bath profile: What*

is it and why should I care? Ottawa, Canada: National Library of Canada. Available: <http://www.nlc-bnc.ca/bath/prof.pdf>.

17. Texas Z39.50 Implementors Group. (1999, April). *Z Texas profile: A Z39.50 profile for library systems applications in Texas, Release 1.0.* Available: <http://www.tsl.state.tx.us/ld/projects/z3950/TZIGProfile99Apr20.htm>.
18. For background on the Z Texas Profile, see Moen, William E. (1998a). *Texas Z: The Texas Z39.50 requirements and specifications project. A discussion paper*. Prepared for the Texas State Library and Archives Commission.
Available: <http://www.unt.edu/wmoen/Z3950/TexasZDPAug98.htm>.
19. See for example, Lunau, Carrol. (1998, June). *Virtual Canadian union catalogue pilot project: Final report.* Ottawa: National Library of Canada. Available: <http://www.nlc-bnc.ca/resource/vcuc/vcfinrep.pdf>.
20. Lynch, Clifford A. (1997, Winter). Building the infrastructure of resource sharing: Union catalogs, distributed search, and cross-database linkage. *Library Trends, 45*(3): p. 449.
21. Coyle, Karen. (2000, March). The virtual union catalog: A comparative study. *DLib Magazine, 6*(3). Available: <http://www.dlib.org/dlib/march00/coyle/03coyle.html>.
22. Hammer, Sebastian and Favaro, John. (1996, March). Z39.50 and the world wide web. *DLib Magazine*. Available: <http://www.dlib.org/dlib/march96/briefings/03indexdata.htm>.
23. Moen, William E. (1998, April). Accessing distributed cultural heritage information. *Communications of the ACM, 41*(4): 45-48. See also the CIMI website at: <http://www.cimi.org>.
24. See Dublin Core Metadata Initiative website at: <http://purl.org/DC/>.

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

- Conference Home Page
- What's new
- Greetings from the Director for Cataloging
- Topical discussion groups
- NAS study and 2 articles from the LC staff *Gazette*
- Conference program
- Speakers, commentators, and papers
- Conference sponsors
- Conference discussion list
- Logistical information for conference participants
- Conference Organizing Team

Barbara B. Tillett, Ph.D.

Director, Integrated Library System Program Office
and
Interim Director for Electronic Resources
Library of Congress
101 Independence Ave., S.E.
Washington, DC 20540-4010

Authority Control on the Web

About the presenter:

Dr. Tillett is currently the Director of the Integrated Library System (ILS) Program at the Library of Congress (LC) that successfully installed a new commercial Integrated Library System for LC on time and on budget by Oct. 1, 1999. For that accomplishment, she received LC's highest honor, the Distinguished Service Award. She also continues as Chief of the Cataloging Policy and Support Office (CPSO), a division of about 60 people that is responsible for various authoritative cataloging tools at LC. She is LC's representative on the Joint Steering Committee for Revision of AACR (JSC). In addition, she is the Interim Director for Electronic Resources, serving to coordinate various initiatives related to processing and accessing "born digital" materials and providing bibliographic control for electronic resources. Tillett has a bachelor's degree in mathematics and master's and Ph.D. degrees in library and information science. Her former positions have included: Head of the Catalog Dept., University of California, San Diego; Director for Technical Services, Scripps Institution of Oceanography; OCLC System Coordinator for the University of California, San Diego; Reference Librarian in science, technology, and medical reference at Hamilton Library, University of Hawaii; and bibliographic analyst and programmer for the Tsunami Document Retrieval System, Hawaii Institute of Geophysics, University of Hawaii. She has also held many positions as teacher and consultant on library automation, cataloging, authority control, and library technical operations, for example, serving as consultant on conceptual modeling to the International Federation of Library Associations (IFLA) Study Group on the Functional Requirements of the Bibliographic Record.

Cataloging Directorate Home Page

Library of Congress Home Page

She also chaired the IFLA working group on defining the minimal set of data elements needed in computer-based, shared, international resource authority records and currently serves on the follow-on IFLA working group to produce functional requirements for authority records. In addition, she serves on two other IFLA working groups on the revision of "Form and Structure of Corporate Headings" and on the revision of "Guidelines for Authority References and Entries," and chairs the IFLA Section on Cataloguing.

Tillett has been active in ALA throughout her 30 years as a librarian, including founding the Authority Control Interest Group in 1984, being chair of the ALCTS Cataloging and Classification Section, and serving on several editorial and review boards for such publications as *Library Resources and Technical Services*, *College and Research Libraries*, and the *ACRL Publications in Librarianship*. She continues to serve on the editorial board for *Cataloguing & Classification Quarterly*. Her publications have focused on cataloging theory and practice, authority control, and library automation. Her dissertation on bibliographic relationships has been a source for conceptual designs for future computer-based systems for bibliographic control.

Full text of paper is available

Summary:

The addition of library catalogs to the mix of information being searched on the Web will open up the Web to focused, topical collections and resources held in and made accessible through the world's libraries. Catalogs have a basic syndetic structure that facilitates finding and gathering together of those resources in whatever media. Authority control enables "precision and recall," which are lacking from today's Web searches. Authority control provides precision to retrieve only those records or items of interest, and the syndetic structure of authority control's cross references assures recall of all the relevant materials, as well as navigation to reach bibliographically related materials.

Explorations to provide interoperability across multiple authority files, to link and provide switching for displays of authorized headings on an international scale, are underway within the International Federation of Library Associations (IFLA). The combinations of Unicode and new technologies are opening up access to all scripts and all languages. Crosswalks, like those provided in OCLC's CORC (Cooperative Online Resource Catalog) project, link Dublin Core (DC) metadata and cataloging rule-based records in MARC and other formats with XML and other communication structures, and expand the opportunities for contributing authority records to an international pool. Standards and agreements are emerging, like a DC for Authorities and the basic data elements recommended in the International Federation of Library Associations (IFLA) "Minimal Level Authority Record."

Other explorations into the use of a standard number for bibliographic entities, as those proposed in the 1970s and more recently by ICA, IFLA, INDECS, and others, may have passed their time of usefulness. Given today's technologies with hyperlinks, URLs, and other mechanisms to connect records and identify and display content, there may be better ways to link, navigate, and display authorized headings.

A pool of authority records for bibliographic entities (persons, corporate bodies, works/expressions, concepts objects, events, and places) to use on the Internet is of interest not only to libraries and their users but also to publishers, copyright and rights management organizations, museums, and archives. We will explore how this all might actually work. Authority control remains the most expensive part of cataloging, but through cooperative efforts like NACO, SACO, and IFLA initiatives, the research done in one library can be shared internationally to lower the cost.

Library of Congress
May 9, 2000
Comments: lcweb@loc.gov

Authority Control on the Web

Barbara B. Tillett

Final version

I won't repeat all the literature reviews (1) and list all of the articles on authority control of this past century, since we all have read them and discussed them. I will instead focus attention on how the authority control performed by libraries can help the Web and suggest some next steps in making this tremendous resource of authority records available and used internationally.

For over a quarter of a century we have been explaining within the library world the virtues of authority control in catalogs, bibliographies, finding aids, and other bibliographic lists to improve the precision of searches and to provide collocation. With the advent of the Internet, and the work of Dublin Core to involve non-library folks in the discussions, the word is out that the libraries may have something there. Queries from intellectual property rights management groups, archives, museums, Web search engine companies, and corporations - some of them are starting to realize they don't have to reinvent the wheel or in this case international authority control. Let's help them achieve it!

THE WEB

The Web is chaotic. Many users get something back and think that's good enough, but may not realize the exact item they need wasn't retrieved. It may be because they tried an author's name and he/she used a pseudonym on the piece they want; or the corporate body they were trying to find changed names or used an acronym that they forgot to search by; or the editors of that famous work didn't use the well known title when it was most recently published. These and many other bibliographic variations cause searches to fail or to retrieve incomplete and sometimes misleading information.

Due to limits on the scope of cataloging, library catalogs don't give you the articles you want and often miss providing access to individual works in a collection or the contents of a compilation or conference proceedings. Authority control won't help that problem, but will help assure getting all the works that were attributed to a particular bibliographic identity (in the Anglo-American cataloging tradition).

When we add library catalogs to the mix of online resources on the Web, we introduce controlled vocabularies for subjects, names (persons, corporate bodies, conferences), and titles. Online catalogs can now serve as gateways to online resources and vice versa. For example, we now provide hypertext links from bibliographic and soon authority records to resources available on the Web. By clicking on the link in the bibliographic or authority record, we launch an Internet connection to the online resource, which may connect to the full text document described in the bibliographic record, or a finding aid cited in the bibliographic record, or perhaps a biographical entry in a reference tool cited in an authority record for a person.

We could also have links on the Web to our catalogs from reference tools or online documents. Those links could allow a researcher to connect from another online tool directly to one or more user-selected online library catalogs to find works by and about the person or corporate body or to topical searches in that library or libraries. We already have this sort of capability in some systems where a user connected to abstracting and indexing services finds they are linked to the resources of library online catalogs that include holdings and location information on where to find specific issues or even to the online full-text article itself. I don't expect abstracting and indexing services to start using authority control for personal or corporate names - that battle has been fought for over a century - but we can help users once they are in the realm of our online catalogs to filter their search results to get what they want and not a lot of extraneous garbage. This may be through authority records that provide access and help distinguish between similar names or provide links from pseudonyms or other variant forms of name or subject terms. (2)

PRECISION AND RECALL

The addition of library catalogs to the mix of information being searched on the Web will open up the Web to focused, topical collections and resources held in and made accessible through the world's libraries. Catalogs have a basic syndetic structure that facilitates finding and gathering together of those resources in whatever media. Authority control enables "precision and recall," which are lacking from today's Web searches. Authority control provides precision to retrieve only those records or items of interest, and the syndetic structure of authority control's cross references assures recall of all the relevant materials, as well as navigation to reach bibliographically related materials. It cannot be stressed enough that this feature of online catalogs adds tremendous value to the user's search and retrieval process. No more wading through tens of thousands of retrieved and computer ranked results for anything close to what we asked for, unless we want to. Let's give users the option for more precise searching, if they want it.

CONNECTING INTERNATIONAL AUTHORITY FILES

From January 1995 through December 1997, the European Commission funded the AUTHOR Project within CoBRA (Computerised Bibliographic Record Actions) to explore the international exchange and re-use of authority records for personal and corporate names. Five national bibliographic agencies participated in a prototype online authority file:

the Bibliothèque nationale de France (Project manager),
the British Library,
the Koninklijke Bibliotheek Albert 1 in Belgium,
the Biblioteca Nacional of Spain, and
the Biblioteca Nacional de Lisboa in Portugal.

Project AUTHOR converted a sample of about 100,000 authority records for a selected set of personal and corporate author names (all names beginning with the letter O or the letter T and a pre-defined set of names of persons and corporate bodies), plus additional records from each national authority file. The USEMARCON universal converter was used to convert authority records to UNIMARC for the prototype database. The database was then accessible using Z39.50 protocol via the Web.

The challenge was that each library has its own language, cataloging rules, bibliographic record format, and local system for its online authority file. There were 5 cataloging languages: Dutch, English, French, Portuguese, and Spanish, plus a bilingual catalog in French and Dutch. There were 5 cataloging rules: AACR2(UK) and national standards for Belgium, France, Portugal, and Spain. There were 5 MARC formats: KBRMARC (Belgium), INTERMARC (France), BLMARC (UK), UNIMARC (Portugal), and IBERMARC (Spain). There were 4 local systems: GEAC (France and Portugal), VUBIS (Belgium), WLN (UK), and ARIADNA (Spain). These factors naturally presented interesting obstacles in sharing authority information.

In their report on findings (3), Sonia Zillhardt and Fran‡oise Bourdon noted that the study revealed different practices and rules for making authority records for specific entities. Although the similarities in rules and practices were great, some obvious differences were apparent. For example, not all of the libraries consider the **names of conferences** as candidates for authority control (Spain, Portugal, and Belgium); or when conferences are included, they are considered corporate names (UK) rather than a separate category (France). The French distinguish between **territorial names** as separate from corporate names, unlike the other libraries. The **use of general explanatory reference records and reference entry records** was not present in this prototype, and indeed are not present in any of the authority files managed by the AUTHOR project partners.

Other differences involve the various **MARC formats** and transliteration practices. The various MARC formats have different elements and tag them differently, for example, in France and Belgium they include nationality of the person or corporate body, but that data element may be just buried in a note if present at all in the UK, Portugal, or Spain. Then there is the **single versus multiple linked record** dilemma for **parallel authorized forms for the same entity** in different languages or scripts. In Belgium, they create a single authority record with the French and Dutch parallel authorized forms of name, and such records were turned into two linked records when converted to UNIMARC for the project. There were also the obvious differences in **transliteration schemes** used by the different libraries.

Earlier the IFLA Section on Cataloguing pointed out some of these same problems when linking single-language and/or multi-language name authority files:

- differences due to language dependent qualifiers or geographical jurisdictions prescribed by cataloging rules
  Elisabetta I d'Inghilterra
  Elizabeth 1, reine d'Angleterre

Elizabeth, Königin v. England

Hungary
Hungria
Magyarország

- different practices with regard to abbreviations
- different transliteration and romanization schemes
- different practices for word division from romanized forms
- differences in filing order, such as treatment of non-filing words, but even more for different scripts
- different MARC (or other communication) formats that have different data elements and the subset of problems, that these formats follow different conventions for codes, such as codes for names of languages
- differences in spelling practices even for the same language (such as US versus most other Anglo-countries -e.g., cataloguing versus cataloging) that cause retrieval problems. (4)

There are also problems in different cataloging codes that traditionally recognize **different entities** as authors and hence providing authority records for those entities. For example, AACR2 recognizes names of ships as "authors" of the ship's logs and recognizes events as entries for publications resulting from that event, while most other cataloging rules do not make this allowance, but instead may include such access as an added entry, if at all. So an authority record in one authority file may not have a counterpart in another national authority file, simply because it is not recognized by the cataloging rules as being eligible. Or there may be differences in the hierarchical levels used by different cataloging rules to represent an entity - such as the conference proceedings of a corporate entity where the AACR2 places conferences as a subheading under the name of the corporate body to group them together. This is a device that collocates the works of that corporate body, but other rules, such as the German RAK (Regeln f•r die alphabetische Katalogisierung) would enter the name of the conference itself or use title entry, not creating the cataloger's corporate heading that AACR2 prescribes. The result is no matches when comparing authority records from one authority file to another.

Another experiment with multiple authority files is being proposed within IFLA (International Federation of Library Associations and Institutions), and several groups have already started work towards creating a virtual international authority file. Unlike the AUTHOR project that created a UNIMARC database of exchanged records from various national authority files, the IFLA project would link existing online authority files through a Z39.50 simultaneous search of the identified national authority files. It would explore ways to provide interoperability across multiple authority files, to link authority records for the same entity through existing record numbers, and to provide switching for displays of authorized headings on an international scale. As a first step, the IFLA UBCIM Working Group on MLAR (Minimal Level Authority Records) and the ISADN (International Standard Authority Data Number) reported its recommendations on the mandatory minimal set of data elements that should be present in all authority records to facilitate international exchange or use. (5) A follow-on group within IFLA, FRANAR (Functional Requirements and Numbering for Authority Records), now is exploring the numbering and functional requirements for authority records. The IFLA Section on Cataloguing Working

Group on FSCH (Forma and Structure of Corporate Headings) is exploring the structures and forms of corporate names to inform developers of future systems and the development of the virtual international authority file. Great benefit may be gained from sharing authority information on an international level. Work continues in this area, and I'll have more to report after the IFLA Conference in Jerusalem in August 2000.

DIGITAL ENVIRONMENT AND METADATA

Crosswalks, like those provided in CORC, link Dublin Core metadata and cataloging rule-based records in MARC and other formats with XML and other communication structures, and expand the opportunities for contributing authority records to an international pool. Standards and agreements are emerging, like a Dublin Core for Authorities (work at the Deutsche Bibliothek) and the basic mandatory data elements recommended in IFLA's "Minimal Level Authority Record," as noted above. [To be expanded]

MULTIPLE SCRIPTS

The combination of Unicode and new technologies are opening up access to all scripts and all languages. Many libraries used to create handwritten catalog entries in book catalogs or old card catalogs and could write in original scripts when transcribing title page information. Even the early printed cards from the Library of Congress included beautiful scripts in the description with accompanying "filing title" information for the transliterated forms of the titles to make it possible to integrate these records into roman alphabet card catalogs. With the early online cataloging tools that multiple script capability was abandoned, because the technology could not handle it. Later the RLIN and OCLC capabilities for selected scripts appeared and now we see on the horizon the potential of using Unicode to present all scripts for all languages in bibliographic and authority records. Within another year this will be a reality in several online systems, opening up the technical capability.

Such possibilities also open up the possibilities of sharing information on a global scale and experiments are already underway. One such experiment is this year's progress among the Hong Kong consortium or research libraries to provide authority records with both Chinese authorized headings in Chinese script and parallel authorized headings from the Library of Congress in the roman alphabet, allowing access from either form.

SWITCHING FOR DISPLAYS

This also gets to a point I've been pushing for a long time - that of "access control" instead of "authority control." I still haven't found another term to use for this concept, but the idea is to control collocation, so the library or the user can select the form of the controlled heading they want to see - the system could switch the display to the chosen form or a default form set by the library. Authority control pulls together all the various forms and relates entities in a way that leads the user to the desired materials and provides a big picture of what is available. With "access control" the same underlying authority records provide control, but the display form is user-selected. In the international context, users may prefer to see headings in their own script, may want to see the names of works in their own language or the names of corporate bodies in a well-known form that may not follow any cataloging rules. Computers let us do this sort of thing through default displays for a given library catalog or a user-selected choice, perhaps recorded in their own computer "client" with an intelligent system making the switch for them before displaying records or entries.

This switching can be accomplished in many ways, such as through using only a number or other identifier for the entity that links to the authority record to display the chosen form. For now that is a single form prescribed by the library, but could also be a form in the users' own language or script preference. Many systems include the authorized form of the name as a text string and may have an associated authority record number for the entity represented by the text string. Through either the text string or the record number link, one can navigate to associated authority records in different countries with different languages and cataloging rules to display their chosen form. This concept is being explored in IFLA.

ISADN (International Standard Authority Data Number)

In 1982 Nancy Williamson predicted catalogs by 2006 would have invisible links of variant forms of names to retrieve all the bibliographic works of a particular person. (6) Many agree this would be lovely, but what would be behind those invisible links to make it work? Some have suggested a standard number.

During the late 1970's an IFLA Working Group led by Tom Delsey suggested establishing an International Standard Authority Number (ISAN) and described the organizational structure for controlling such numbers and their assignment and maintenance. Delsey recognized the practical aspects of administering such a number was "far from simple." (7) The idea of the ISADN (International Standard Authority Data Number) was reiterated in the Guidelines for Authority and Reference Entries published by IFLA in 1984.(8) Unfortunately the cost of an international organization to manage such a system was prohibitive and technology had not yet advanced to a point to assist such international control, so the idea fell by the wayside.

A model put forward by Snyman and Jansen van Rensburg suggested using an International Standard Author Number (ISAN) (9), which they later label as "INSAN." (10) Despite unfortunate problems with their historical facts and citations to earlier work in this area, Snyman and van Rensburg offer the same solution of a single number used universally. Their number contains 18 characters:

the first two alphabetic characters to identify the agency responsible for issuing the author number, the next two alphabetic characters identifying the nationality of the author (a big problem for the United States where we tend to catalog materials for authors worldwide and not just for our own country), the next 3 alphabetic characters to identify the language typically used in the author's original publications (also problematic for the current world's authors), the next 4 numeric characters for the year of issue of the number, the next 6 numbers to be a serial number assigned incrementally for the INSAN - allows for a million new "authors" per year per agency), and a final check digit at the end.

There have been many calls for the use of a standard number for bibliographic entities, such as those proposed in the 1970's and more recently by ICA, IFLA, , and others. But has such a single number approach passed its time in terms of what we can now technically accomplish? Given today's technologies with hyperlinks, URL's, and other mechanisms to connect records and identify and display content there may be better ways to link, navigate, and display authorized headings.

The simplicity and elegance of having a single number to universally control the names for persons, corporate bodies, and works persists to this day. It is attractive to those dealing with copyright and other intellectual property rights, to archives and libraries and museums wishing to share the cost of bibliographic and authority control. The IFLA UBCIM Working Group on Minimal Level Authority Records and ISADN in its final report, "Mandatory Data Elements for Shared Authority Records" in 1998 stated that such numbers may not be needed if one used instead the existing authority record control numbers and linked them across the authority files of the major national bibliographic agencies. Despite this recommendation, some members of IFLA itself persisted in calling for an ISADN (International Standard Authority Data number). So another IFLA UBCIM Working Group is now in progress looking yet again at the functional requirements and numbering for authority records (FRANAR).

AUTHORITY RECORD RESOURCES

A pool of authority records for bibliographic entities (persons, corporate bodies, works/expressions, concepts, objects, events, and places) to use on the Internet is of interest not only to libraries and their users but also to publishers, copyright and rights management organizations, museums, and archives. We already have several major authority files created by national bibliographic agencies, such as the Library of Congress and the national libraries of Belgium, France, Germany, Italy, Portugal, and Spain, to name a few. This wealth of information provides a huge resource that hold great potential for enabling the controlled access of the future on a global scale across many applications for libraries, intellectual property rights, archives, museums, etc.

MODEL

Let's explore one scenario for how this all might actually work for name and title authority data. Taking a practical approach to use what we already have rather than to establish a complex system to create yet another control number, let's look at one model.

There are multiple objectives:

- to facilitate sharing of authority information to reduce the cataloging costs among libraries and other users of such data (such as archives, museums, and agencies for intellectual property rights management) - this is the current driving objective - with the corollary

- to simplify the creation and maintenance of authority records internationally, with the ultimate goal
- to enable users to access bibliographic information through controlled access that lets them select the form of names they prefer (either to select the script they prefer, the language they prefer, or even the form/structure of the name they prefer) or to display a default established by the library of choice.
- This is turn can facilitate links between resources: connecting the user to publications, articles, materials, and objects including all those that are digitally available or can be ordered/requested by the user, expanding the online catalog to a more comprehensive gateway to knowledge.

All very grand and wonderful, but where to begin?

We have the existing authority files from major national bibliographic agencies and IFLA can maintain a list of those agencies that would be willing to share their authority records on the Web.

We know the composition of those authority records from earlier IFLA studies and can map the mandatory data elements for a Z39.50 profile. [recommendation: have LC establish such a profile]

How to assure pulling up all existing authority records for the same entity when forms may vary from file to file? Rather than creating a worldwide system for assigning an ISADN, we can use the existing authority record control numbers to provide a link, and to make display easier, it may be useful to also include the text string for the authorized form for the name.

That text string and record identifier from another nation's authority file could be used to switch the form of headings in shared bibliographic records either when cataloging or when displaying the records to users.

As noted in the IFLA recommendations on mandatory data elements in shared authority records, such authority records should have the text strings for authorized forms and variant forms of name and a record of related entities, as well as the number for the entity. (There are 19 mandatory data elements prescribed by IFLA and another 3 elements that are highly recommended.) (11) That entity identifier may be reflected in the record number for the bibliographic identity of the entity, such as LC's control number (LCCN). That number needs to uniquely identify the record for the entity (or bibliographic identity in AACR2 terms) and by extension it can be used to identify the entity/bibliographic identity itself.

As a quick aside, note the distinction between **authority records** - a device to record decisions, used for maintenance and display of a chosen authorized form and links (references) from variant forms of name for a given bibliographic identity/entity and links with names for related entities (see also references) and **authority entries or authorized headings** - the chosen controlled form of the heading used as the access point or the display form for the name of an entity.

The internal workings of an online system can store text strings, numbers, codes, or other mechanisms to then display the authorized heading to the user. How a system accomplishes this should be transparent to the user. The system could also display any chosen form the user requests or a default form chosen by the agency presenting the information to the user. Currently, we use a default form that is specially tagged in an authority record as the authorized form according to our cataloging rules.

So back to our model.

Original Cataloging
Scenario 1 - match found for same entity

1. Cataloger A begins cataloging a new item (creating a bibliographic record) and identifies a name (personal, corporate, conference, or uniform title). Cataloger A's online system automatically checks the local authority file and if not found also checks the virtual international authority file (automatically launching a Z39.50 connection behind the scenes) to see if that name has already been established somewhere in the world. Hopefully future online connections will be much faster and more reliable than they are today!
2. Let's say the name was established by Cataloger X across the world and the record was found and displayed to Cataloger A.
3. Cataloger A confirms it is the same entity.
4. If the Cataloger X authority record can be used as is, Cataloger A lets the system know it is ok.
5. The record is automatically added to the local authority file with a system generated local authority record control number, preserving the control number and text string found in Cataloger X's original authority record.

Variation on Scenario 1 - the international authority file finds two or more matches and displays them to cataloger A who then can select the one that is closest to their needs (more complete or matches the language and /or cataloging rules used by Cataloger A's library).

Original Cataloging
Scenario 2 - match found for same entity but needs editing
Same as Scenario 1 except for step 4 where Cataloger A decides to edit the existing authority record to meet the cataloging rules and practices of Cataloger A's library. Cataloger A then lets the system know the record is ready and
5. is the same - The record is automatically added to the local authority file with a system generated local authority record control number, preserving the control number and text string found in Cataloger X's original authority record.

Original Cataloging
Scenario 3 - match found but too time consuming to edit
Same as Scenario 1 except for step 4 where Cataloger A determines editing would be too time

consuming and has the system generate an automatic authority record that is then edited as needed.
5. Cataloger A confirms and tells the local system this is the same entity.
6. Same as scenario 1, step 5 - The record is automatically added to the local authority file with a system generated local authority record control number, preserving the control number and text string found in Cataloger X's original authority record.

Original Cataloging
Scenario 4 - match determined to be for different entity
Same as scenario 1 except Step3 where Cataloger A determines the found authority record from Cataloger X is not for the same entity.
4. The local system creates an automatic authority record that can then be used or edited and added to the local authority file.

Original Cataloging
Scenario 5 - no match found in local or international authority file
The local system creates an automatic authority record that can then be used or edited and added to the local authority file.

Original or Copy Cataloging
Scenario 6 - match in local authority file without an internationally linked authority record
Let's say now Cataloger B in the same library keys in the heading for the same entity on an original record for another item or the heading appears on a record from copy cataloging. The local system finds the matching authority record and alerts the Cataloger B that the heading is already established.

Note that the cataloger could choose to launch a check against the international authority file at this point if desired for future links.

Copy Cataloging
Scenario 7 - match in local authority file with internationally linked authority record
If cataloger B is doing copy cataloging and brought in a record from another country, that used a different form for its authorized heading and the system discovers the heading in the local authority file, notices the text string matches a parallel authorized form from that other country (preserved when the international authority file record was captured) and either automatically switches the form in the bibliographic record to match the Cataloger C's authorized form, or displays that form (or the users chosen form) on the fly when the record is presented to a user. This display capability actually could apply to any of the scenarios for alternate forms for the name found in the authority record.

Copy Cataloging
Scenario 8 - no match in local authority file
Catalogerr C brings in a copy cataloging record from another country with no matches in the local authority file, so the system launches the search of the virtual international authority file and displays any matches (either on a reference or an authorized form or near matches). If a match, then we are back to the same process as with Cataloger A - either use the record as is, edit it, or create a new authority record,

linking when it's the same entity. If no match, then the local system automatically generates a base authority record that the cataloger can use or edit as needed.

Are you getting the idea? And you may have other suggestions for how this could play out.

ONE SIZE DOES NOT FIT ALL

[Why need cross references that go with the rules and chosen authorized heading..and how the alternate authorized form from other cataloging rules can be used as either another variant form (x ref, see from)) or a related heading (xx ref, see also from). - to be expanded]

Why not create one giant authority file that combines all the variant forms from all the authority files and lets the user decide which form to display? Cross references follow rules for a given catalog's syndetic structure. One cannot just combine all the references from different authority records created from different cataloging rules and principles and have it work elegantly. Differences in display order and filing rules, rules for additions to and omissions from names - all combine to destroy syndetic structures. Users would find themselves buried in variations.

SUBJECT AUTHORITY

I've focused mostly on sharing of name and title authority information, but there is the whole universe of subject authority control and efforts to link various subject heading schemes, thesauri, and subject classification systems. Experts, such as Karen Markey Drabenstott in particular, have pointed to ways to improve subject searching on the Web and much work is still needed in this field. [citations to be added] Gail Hodge also recently suggested using knowledge organization systems that include authority files, glossaries, dictionaries, gazetteers of place names, classification schemes, etc. to help structure digital libraries and this can be extended to online information on the Web. (12.)

FUTURE

Authority control remains the most expensive part of cataloging, but through cooperative efforts like NACO, SACO, and IFLA initiatives, the research done in one library can be shared internationally to lower the cost. We have a wonderful opportunity to really make this work. More prototyping is rumored to be going on in Europe and Chinese libraries are beginning to make links across national authority files. Let's do it.

---

1. For the curious, some examples are Arlene Taylor's "Research and Theoretical Considerations in Authority Control: in Tillett, Barbara B. Authority Control in the Online Environment. Haworth Press, 1989 (also published as Cataloging & Classification Quarterly, v.9, no. 3, 1989, p.29-56. Larry Auld's "Authority Control: An Eighty-Year Review," Library Resources & Technical

Services, (Oct./Dec. 1982), v. 26, p. 319-330. Barbara Tillett's "Automated Authority Files and Authority Control: A Survey of the Literature," seminar paper, Graduate School of Library and Information Science, University of California, Los Angeles, June 1982; with corrections and additions, October 1982.

2. I recently came across an article that describes what I was already exploring with Oxford University Press, namely using authority records as a link with biographical information. The Web article provides an interesting set of suggestions to "improve the organization of digital libraries and facilitate access to their content" (abstract) - seeGail Hodge "Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files," CLIR Publications & Resources, pub91 (April 2000) (Available on the Web as: http://www.clir.org/pubs/reports/pub91)
3. Zillhardt, Sonia and Françoise Bourdon. AUTHOR Project : Transnational Application of national Name Authority Files, Library Project PROLIB/COBRA-AUTHOR 10174, Final report. Paris : Bibliothèque nationale de France, 1998 (available from the authors).
4. Murtomaa, Eeva and Eugenie Greig with help of Joan Aliprand. "Problems and Prospects of linking Various Single-Language and/or Multi-language name Authority Files," International Cataloguing and Bibliographic Control, v. 23, no. 3 (July/Sept. 1994), p. 55-58
5. IFLA Working Group on MLAR and ISADN. Mandatory Data Elements for Internationally Shared Resource Authority Records : Report of the IFLA UBCIM Working Group on Minimal Level Authority Records and ISADN". [Frankfurt]: International Federation of Library Associations and Institutions, Universal Bibliographic Control and International MARC Programme, 1998.
6. Williamson, Nancy J. "Is there a Catalog in Your Future? Access to Information in the Year 2006," Library Resources & Technical Services, v.26 (April 1982): p, 122-135
7. Delsey, Tom. "Authority Control in an International Context." In: Tillett, Barbara B. Authority Control in the Online Environment : Considerations and Practices. New York: Haworth Press, 1989., p. 25.
8. Guidelines for Authority and Reference Entries, recommended by the Working Group on and International Authority System, approved by the Standing Committee of the IFLA Section on Cataloguing and the IFLA Section on International Technology. London: IFLA International Office for UBC, 1984.
9. Snyman, M. M. M. [and] M. Jansen van Rensburg. "Reengineering name authority control," The Electronic Library, v. 17, no. 5 (Oct. 1999), p. 313-322.
10. Snyman, M. M. M. [and] M. Jansen van Rensburg. "Revolutionizing name authority control," Digital Libraries, San Antonio, TX, ACM, 2000, p. 185-194.
11. Op. cit. IFLA Working Group on MLAR and ISADN. Mandatory Data Elements for Internationally Shared Resource Authority Records, p. 3-6.
12. Hodge, Gail. "Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files," CLIR Publications & Resources, pub91 (April 2000) (Available on the Web as: http://www.clir.org/pubs/reports/pub91)

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Ann Huthwaite
Bibliographic Services Manager
Queensland University of Technology
Library
Kelvin Grove Campus
Victoria Park Rd.
Kelvin Grove Qld 4059
Australia

AACR2 and Its Place in the Digital World: Near-term Revisions and Long-term Direction

About the presenter:

Ann Huthwaite is currently the Library Resource Services Manager at the Queensland University of Technology Library, where she is responsible for the cataloguing and acquisitions functions. She has been the Australian representative on the Joint Steering Committee for Revision of AACR (JSC) since 1994, and was appointed Chair of the Committee in 1999.

Ann has a long involvement with cataloguing in Australia. She has been a member of the Australian Committee on Cataloguing since 1992, and was that committee's representative on the ABN Standards Committee. She was a joint editor of *Cataloguing Australia* for several years, and convened the 13th National Cataloguing Conference in 1999. She has served on the executive of the Queensland Group of the ALIA Cataloguers' Section since 1987, generally as President of the Group.

Prior to her appointment at the Queensland University of Technology as the Cataloguing Librarian in 1989, Ann worked in various positions at the State Library of Queensland.

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

Ann holds a Bachelor of Arts degree and graduate diplomas in education and librarianship. She also holds a Master in Applied Science (Information Studies) from Charles Sturt University. The focus of her research for this degree was user interaction with the catalogue.

Full text of paper is available

Summary:

The context in which cataloguing operates has changed significantly since AACR2 was first published. We have seen the emergence of new media and new modes of publication. Electronic documents are less stable and more difficult to define than their print counterparts. The ready availability of networked resources on the Internet has changed the way in which users obtain and use information. More stakeholders are involved in the provision of access to bibliographic resources. The ability of the current rules in AACR2 to adequately describe electronic resources has been called into question. The emphasis on the item in hand--the physical object--is considered inappropriate for cataloguing remote access electronic resources. The related class of materials concept has also been shown to be flawed. The proliferation of records for the same work is becoming confusing for users, particularly for serials published in different formats. The Joint Steering Committee for Revision of AACR (JSC) is acutely aware of the concerns expressed by the cataloguing community about the adequacy of the existing rules, and significant work has taken place in recent years to address these concerns. Work undertaken includes the organization in 1997 of the International Conference on Principles and Future Development of AACR. The relevancy of the rules in the online environment was a major focus. Principal outcomes included the commissioning of three reports: a logical analysis of the rules by Tom Delsey, using a data modeling technique; a report on seriality; and a proposal to revise Rule 0.24 to advance the primacy of intellectual content over physical format. Several initiatives are being pursued as a result, including a major revision of Chapter 12, a revision of Rule 0.24, an expanded introduction, and a new appendix defining major and minor changes. At the same time, a major revision of Chapter 9 has been in progress, to bring the rules into closer alignment with the International Standard for Bibliographic Description for Electronic Resources (ISBD(ER).

This paper will review progress on these developments, including the outcomes of the JSC meeting to be held in London in September, and will focus on the implications for the cataloguing of electronic resources. JSC is also considering suggestions for the reorganization of Part 1 of AACR2 according to ISBD areas. The first stage of a prototype has been developed to test the feasibility of the proposal. This paper will review progress to date, and will present the advantages and disadvantages of the proposed restructure.

Various possibilities for the long-term direction of AACR will be explored,

bearing in mind that JSC members represent their constituent bodies, and decision-making takes place in a consultative environment. Future changes to the code will be in the hands of the Anglo-American cataloguing community, not a small group of individuals. The paper will also explore the relationship between AACR2 and metadata schemes. It will present the case that both sets of standards have a role to play, and that AACR will continue to be used for electronic resources of lasting value.

Lynne C. Howarth, commentator

Associate Professor and Dean
Faculty of Information and Studies
University of Toronto
140 St. George St.
Toronto, Ontario
MSS 3G6, Canada

B.A. (McMaster),
MLS, Ph.D. (Toronto)

About the commentator:

Lynne Howarth completed her Ph.D in library and information science and was appointed to the Faculty of Information Studies, University of Toronto in 1990, becoming Dean in 1996. Prior to that she worked as Cataloguing Supervisor, then Systems Librarian at North York Public Library, and taught cataloguing and classification at McGill University (Montreal), and principles of information management at Ryerson Polytechnical University. Current teaching and research are focused on the creation and application of bibliographic tools and standards, organization and management of technical services, and knowledge management applications in private and public sector institutions. She was recently awarded a three-year grant from the Social Sciences and Humanities Research Council of Canada to develop a metadata-enabled web search prototype. She is a member of the IFLA Section on Cataloguing and chairs its Working Group on Metadata Schemes.

Full text of commentary is available

Library of Congress
December 21, 2000
Comments: lcweb@loc.gov

AACR2 and Its Place in the Digital World: Near-Term Solutions and Long-Term Direction

Ann Huthwaite
Bibliographic Services Manager
Queensland University of Technology Library
Kelvin Grove Campus
Victoria Park Rd.
Kelvin Grove Qld 4059
Australia

Final version

Preface

As Chair of the Joint Steering Committee for Revision of AACR (JSC) I am conscious that I wear several hats. I wear an official one when I am acting as the spokesperson for the committee. There is an Australian version that I wear when I represent the views of the Australian cataloguing community. Then there is my own hat that I wear when I express my personal views. Throughout most of this paper I wear my JSC hat (as I have been asked to contribute to this conference in that capacity). However, there are times when I give my personal views, and I will make it clear when I have changed hats to do this.

Changes in the bibliographic universe

The universe of information has changed significantly since AACR2 was first published in 1978. It is now hard to remember what it was like before personal computers and the Internet became part of our daily lives. In retrospect, the changes we have all been through represent a true paradigm shift in the way we communicate and distribute information.

In 1978 print was the predominant medium of recorded communication. Formats such as videorecordings and audiocassettes were in use, but we were using them rather self-consciously, almost as adjuncts to the real thing. Since then a variety of new media has emerged, and instead of being just added extras or embellishments they are the real thing. Some new formats are notoriously difficult to pin down and classify. For example, a DVD can be considered either a computer file or a videorecording, depending on the nature of the content. A digital map is both computer file and cartographic material.

Twenty-two years ago there were serials and there were monographs. Loose-leaf publications were problematic, but we were able to squeeze them into the monographic mould. Electronic publishing has unleashed a whole new stable of hybrid beasts. Electronic documents are unstable; they can transform into different versions of the original or completely different creations. A book or a videorecording has a visible boundary; we can see where it begins and ends. Documents on the Internet are not so clearly defined.

Users in this new environment have completely different expectations. Although they are still seeking information to fulfil their needs, they expect it now (and in full text). They are more sophisticated in their searching techniques, and there is a great variety of different approaches. At the same time, the ready availability of information has created a new set of problems associated with information overload.

There are now more players in the business of information storage and retrieval. Of course AACR2 has never been the sole system for providing intellectual access to bibliographic resources - library cataloguing encompasses only a subset of the information universe - but now experts are emerging in many other domains and disciplines. Metadata developers, for example, are now debating the very issues that cataloguers have been dealing with for generations.

At the same time that this revolution has occurred there has been growing pressure on publicly funded institutions to reduce costs. Libraries throughout the world have been cutting back on expenditure and services. It is ironic - and perhaps tragic - that at a time when our profession's most creative minds should be applied to studying the implications of the changed environment that we are forced to focus on local and short-term issues.

Perceived shortcomings of AACR2

The rules in AACR2 were intended to be used for any type of material, including electronic resources. Over the last few years this underlying assumption has been challenged. The emphasis on the item in hand - embodied in the method of procedure stated in rule 0.24 in the introduction to part I - is seen to be inappropriate for cataloguing remote access electronic resources. In its current form, this rule states: "It is a cardinal principle of the use of part I that the description of a physical item should be based in the first instance on the chapter dealing with the class of materials to which that item belongs." Is it logical, or indeed possible, to apply this method of procedure when there is no physical item in hand - when the bibliographic entity exists in digital form on a remote computer?

The class of materials concept also appears to be breaking down. Some types of new media do not fit neatly into a given class of material, and may display characteristics of more than one class. The work conducted recently by Tom Delsey has shown that different criteria are applied to assign bibliographic entities to the specified classes, which does not bode well for the code's ability to extend to new and emerging media. The underlying principles must be internally consistent if the code is to expand indefinitely.

There is also concern that the current rules are not sufficiently flexible to adequately describe materials that change over time - a common characteristic of electronic resources. The snapshot approach that has been applied does not work so well for resources that leave little evidence of when the changes took place and what the changes were.

In the past, documents were normally produced in a single format - print - with occasional examples of reproduction in another medium, usually microform. The rules in AACR2 require the starting point for description to be the physical form of the item in hand, not the original or any previous form in which the work has been published. This requirement has led to what has become known as the "multiple versions" problem, where several catalogue records can exist for the same work. The rapid growth in electronic publishing has compounded this problem and is causing real inconvenience for catalogue users. Many libraries have adopted a "single record" approach for cataloguing their journal collections, based more on expediency than on sound principles. As the trend towards parallel print and electronic publication extends to monographs, this problem can only get worse.

Processes for change

The responsibility for ongoing revision of AACR2 rests with the Joint Steering Committee for Revision of AACR (JSC), working in conjunction with the Committee of Principals of AACR. The members of JSC have been acutely aware of the concerns expressed by the cataloguing community about the ability of the existing rules to adequately describe electronic resources, and over the past few years the committee has embarked on an ambitious program of reform.

JSC sometimes comes under fire for its slowness in responding to perceived problems (the word "glacial" has been used to describe its progress). It is by nature a consultative committee, and to a large extent bound by the decisions of its constituent bodies. This is the source of both its strength and its weakness. By seeking wide input it benefits from considered, specialised opinion and has a sound mandate for action. However, it cannot move as quickly as many would like.

In recent years JSC has been proactive in seeking solutions, but has stressed the importance of taking a fundamental, long-term approach rather than applying short-term, "band-aid" measures. In particular, it believes that we must first deeply understand the principles embodied in the existing rules to determine whether they are sufficiently sound and internally consistent to support ongoing change.

International Conference on the Principles and Future Development of AACR

At the JSC meeting in Boulder, Colorado, in 1994, the idea was first mooted to hold an international conference of cataloguing experts to discuss the main issues facing AACR2 and to provide direction to the committee for the ongoing development of the rules. The idea gained momentum, and in 1997 JSC organised the International Conference on the Principles and Future Development of AACR. The conference was held in Toronto, Canada with sixty-four invited participants. Following the presentation of papers at the conference, several discussion groups were formed to discuss the main topics and to recommend further action.

JSC met immediately after the conference to establish a plan to be implemented in conjunction with the Committee of Principals of AACR. A number of items for immediate action were identified. The items with particular relevance to this discussion were:

1. To pursue the recommendation that a data modeling technique be used to provide a logical analysis of the principles and structures that underlie AACR;
2. To create a list of the principles of AACR2;
3. To formalise the recommendations on seriality endorsed during the conference and introduce them into the rule revision process;
4. To solicit a proposal to revise rule 0.24 to advance the discussion on the primacy of intellectual content over physical format (the "content vs carrier" problem).

The following is a report on progress with these items.

Action item 1: data modeling

Tom Delsey was commissioned to undertake a logical analysis of the code using the entity-relationship technique used previously by the IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR). The schema subsequently developed was intended to serve as a tool to assist in an examination of the principles underlying the code. Two reports were prepared for the two parts of AACR2, each accompanied by several recommendations.

The model has revealed a complex underlying structure, with some anomalies and inconsistencies. The concept of class of materials has not stood up well to the analysis, and Delsey has recommended that options for restructuring part I of AACR2 be explored, with one option the use of the General International Standard Bibliographic Description (ISBD(G)) areas of description as the primary organising element. Progress on this recommendation is outlined later in this paper.

Delsey's conclusions about the code's ability to adequately describe "continuing" publications and materials that change over time have strongly influenced the rule revision proposals on seriality (also described later in this paper).

Delsey's analysis of part II of AACR2 calls for a re-examination of some of the more fundamental concepts of the code, such as "authorship," "work," and "edition." The issues are complex, and JSC has decided to move forward in the first instance on a more practical issue, the limitations imposed by the "rule of three." The rationale for this rule has its origins in the card catalogue era, and makes little sense in an online environment. The Australian Committee on Cataloguing is currently working on rule revision proposals to make this limitation an option.

Action item 2: list of principles

JSC has compiled a list of principles based on submissions by its members. Barbara Tillett is doing further work on refining this list, and her report will be discussed by JSC at its September 2000 meeting.

Action item 3: revising AACR2 to accommodate seriality

The paper by Jean Hirons and Crystal Graham, "Issues Related to Seriality," aroused a great deal of interest at the International Conference, and there was consensus that rule revision proposals should be prepared to move their recommendations forward. Jean Hirons was asked by JSC to coordinate the revision process. Hirons has been working closely with the ISBD(S) and ISSN communities to facilitate harmonisation of the three sets of standards.

Rule revision proposals particularly relevant to the cataloguing of electronic resources include the following:

- The extension of the scope of the current chapter on serials to all continuing resources, including integrating resources, such as loose-leaf publications and databases (with the chapter proposed to be called "Continuing Resources" instead of "Serials");
- The inclusion of rules specific to integrating resources;
- The inclusion of rules specific to electronic continuing resources;
- The inclusion of special rules for remote access serials that are not organised in issues and that lack the kind of bibliographic information present in print serials;
- The inclusion of examples relevant to electronic continuing resources.

At the time of writing these proposals had not been discussed by JSC. They are currently being reviewed by the constituencies and will be considered by JSC at its next meeting in September 2000.

Action item 4: content vs carrier

The ALA/ALCTS/CCS Committee on Cataloging: Description and Access (CC:DA) was asked to create a proposal to advance the discussion on the primacy of intellectual content over physical format. This is not a trivial issue, and the report presented to JSC by a CC:DA task force in September 1999 was both detailed and comprehensive.

The task force pointed out that the current rule 0.24 has two main functions. Firstly, it instructs the cataloguer to assign a bibliographic item to a particular class of material, and gives some guidance on how to describe an item when it exhibits characteristics of more than one class. Secondly, it gives indirect guidance on when to create a new bibliographic record; it implies that two items containing identical content (i.e. containing the same expression of the same work) but stored in different physical carriers should have separate records.

The report contained three recommendations. The first proposed a change to the wording of rule 0.24, emphasising the need to bring out all aspects of the item being described. The wording of the proposed revision, with one change, was endorsed by JSC at its March 2000 meeting and will be incorporated in the next revision package.

The second recommendation dealt with the complex issue of format variation, or multiple versions. JSC agreed with the proposal that explicit guidance on when to create new records should be included in the rules; the draft of an appendix containing such guidelines is currently underway. JSC also agreed that further investigation of this vexed issue is needed, and will set up a working group to move this forward. The working group will be asked to consider as a starting point an option developed by the task force that instructs the cataloguer to ignore any mere physical variation or any mere variation in distribution information (i.e. any manifestation variation) in determining when to make a new record. This option would require that the definition of "edition" in AACR2 be revised to be more in conformity with the definition of "expression" in FRBR.

The third recommendation echoed a proposal in the seriality recommendations to include a statement of principles and other information in the introduction to AACR2 to clarify and facilitate the cataloguing process. JSC expects to review a draft of this expanded introduction at its September 2000 meeting.

Alignment of ISBD(ER) with AACR2

The ALA submitted a proposal in 1998 to begin the process of harmonisation of the rules with the recently published International Standard Bibliographic Description for Electronic Resources (ISBD(ER)). The change of the General Material Designation (GMD) from "computer file" to "electronic resource" was seen to be a high priority.

As JSC further considered the proposals, it became clear that this was not going to be a simple matter, and that far more was required than the mere substitution of one term for another. Since that time the ALA proposals have undergone a number of iterations, and JSC hopes to be able to finalise the revisions at its September 2000 meeting. A completely revised Chapter 9 (to be renamed "Electronic Resources") will be the result, together with the revision of a number of associated rules in other chapters.

The revisions have not always followed the ISBD(ER), and in some cases have gone beyond it. The task force decided that complete harmonisation was neither possible nor appropriate in some cases, and subsequent consideration of the proposals by the constituent bodies has resulted in further refinement. Significant areas of change from the current rules include:

- The use of a new GMD ("electronic resource");
- The updating of terminology throughout;
- The expansion and clarification of the scope of the chapter;
- A new definition of chief source (the resource itself instead of the title screen(s));
- The inclusion of references to the new appendix (to give guidance on when to create new records);
- The inclusion of examples appropriate to contemporary electronic resources (particularly in the note area);
- The updating of terms in the Glossary, using ISBD(ER) terminology where appropriate.

At the March 2000 meeting, two new and significant proposals were put forward by the Library of Congress; these are currently under discussion by the constituencies.

Firstly, the Library of Congress proposed that Area 3 (currently the File Characteristics Area in AACR2 and Type and Extent of Resource Area in ISBD(ER)) be removed from Chapter 9, or at least made optional. It does not support the inclusion of the list of designations as given in this area in the ISBD(ER), maintaining that it amounts to little more than a list of genre terms, which would be difficult to keep current. It proposes that the information recorded in this area could be transferred to the note area.

The current rules instruct the cataloguer not to give a physical description (in Area 5) to remote access electronic resources, despite the fact that this area contains information relating to the content of an item. The Library of Congress has proposed a re-examination of the logic of this exclusion, in light of the change of emphasis in the rules from "carrier" to "content."

Another issue yet to be resolved is whether all remote access electronic resources should be considered published. ISBD(ER) has taken this practical approach and the JSC constituent bodies are currently considering whether to follow suit.

Reorganisation of part I of AACR2 according to ISBD areas of description

Tom Delsey recommended that consideration be given to reorganising part I of AACR2 according to ISBD areas of description, and this suggestion has been supported by other groups - in particular those dealing with the issues of seriality and content vs carrier.

JSC has been pursuing the suggestion. The first stage of a prototype has been developed by Bruce Johnson and Bob Ewald from the Library of Congress, using Cataloger's Desktop to rearrange the current rules under each area.

Not all constituent groups support the proposal. The Australian Committee on Cataloguing, for example, is not convinced that the reorganisation would achieve a great deal. It believes that the end result could be complex and unwieldy, and that it would still not address the difficulties associated with cataloguing material that exhibits characteristics of more than one class.

A simple reorganisation of the rules still preserves the class of materials concept. The cataloguer still has to select a predominant class in order to determine chief source and prescribed sources of information, General Material Designation (GMD), and Specific Material Designation (SMD). Cataloguers describing particular types of material may have difficulty locating the relevant rules. A serials cataloguer, for example, would have to go through the rules for each ISBD area to find the rules relevant to serials.

The CC:DA Task Force on Rule 0.24 has pointed out that the GMD remains one of the most intractable problems when considering the content vs carrier issue. It has presented some options for dealing with the problem, including the provision of a table of preference for selection of the GMD, allowing the formulation of compound GMDs, and abandoning the use of GMD altogether. The Library of Congress is currently preparing a discussion paper on this issue for review by the JSC constituent bodies before the September 2000 meeting.

JSC examined the prototype at its March 2000 meeting, and decided that in the interim a more sensible approach may be to consolidate the rules for general description (chapter 1). To achieve this the rules for each type of material would be examined to determine whether they could be generalised and moved to chapter 1.

In its initial response to the Delsey papers the Australian Committee on Cataloguing suggested that consideration be given to exploiting the potential of the electronic version of AACR2. The current electronic version is little more than the print version transferred to electronic form. In an ideal world a cataloguer should be able to re-order and customise the rules according to the needs of the moment.

Additional issues associated with the cataloguing of electronic resources

Some additional problems associated with the cataloguing of electronic resources remain unresolved and will need to be considered by JSC when it is working on the list of principles that underlie the code. The following are some problems that I have identified; there may be others yet to be raised by the cataloguing community or yet to be manifest. As cataloguers continue to gain more experience in the cataloguing of electronic resources other issues will surface, and issues that seem to be a problem now may disappear. It will be an iterative and evolving process, and we would be foolish to think that we are at the end.

Defining the boundaries of electronic documents

In his analysis of part I of AACR2, Tom Delsey pointed out that the current rules normally assume that the entity being described is a physically discrete object (Delsey, 1998, p.29). If this remains central to the logic of the code then it becomes difficult to define the boundaries of a document not defined in physical terms. For example, when cataloguing a Web site the cataloguer has to decide whether the document (or "information package") is the Web site itself or to include documents attached by hypertext links.

Describing electronic documents with presentation variations

An electronic document stored remotely can alter depending on the software used to display it. Not only can there be variation in style but also variation in content. The current rules are based on the assumption that one copy of an item is identical to another.

Describing electronic documents that change over time

Delsey has highlighted the shortcomings of the existing rules when it comes to describing electronic documents that change over time (Delsey, 1998, p. 34-35). The snapshot approach that has worked well enough for print documents is more difficult to apply to electronic documents that may not leave any clues about when changes have occurred and how the content has altered. Two cataloguers describing the same document at different times may give quite different descriptions. Delsey proposes that the code should allow multiple values of an attribute that changes over time, with the problem then being to decide how to represent these multiple values in the description.

JSC's program of work

With so many revisions planned and underway, the task of coordinating JSC's program of work is becoming increasingly complex. Four main areas of change are interdependent and should occur simultaneously: the revision of chapter 9 (computer files/electronic resources); the revision of chapter 12 (serials/continuing resources); the expanded introduction; and, the new appendix. The revisions to chapter 9 are very nearly finalised but those to chapter 12 are not so well advanced. JSC would like to be able to incorporate the revisions into a revision package at the end of the year, but this assumes widespread agreement by the constituent bodies.

JSC members keep in touch by email but are largely constrained by the meeting timetables of the bodies they represent. It must also be borne in mind that all the members have busy working lives and can only commit part-time hours to their JSC responsibilities. Even with the best of intentions it is difficult to introduce major changes quickly. However, there is considerable momentum at present and we do expect that AACR2 will undergo significant change over the next five years.

Long-term direction for AACR

This is where I must take off my JSC hat and don my personal hat. Let us jump ten years hence to see what AACR might look like and what role it might be playing.

Scenario 1: More of the same

Print still dominates the publishing industry. Despite predictions about the demise of the book it is still flourishing. Libraries look much the same as they do today.

AACR also looks much the same but may be AACR4 or AACR5. The electronic version is more widely used but the print version is still popular.

It is still arranged in much the same way, with chapters in part I devoted to different types of material. There are difficulties with the class of materials concept but cataloguers are making do. They would like it to be easier but accept that no better alternative is easily found. Most of the time they get it right and most of the time users find what they are looking for in library catalogues.

The MARC format is still in use - creaking at the edges perhaps, but the library industry does not have the resources to invest in developing a better medium.

Scenario 2: A hybrid universe

Print and other tangible formats are still widely used, but electronic publishing is starting to dominate the industry. The technology and usability of the e-book have vastly improved, and young people in particular have abandoned print. Book stacks are disappearing from libraries and being replaced by computer terminals.

AACR is published only in electronic form. It is no longer called "AACR" but something like "International metadata standards for information centres." It is a fine example of an electronic manual - flexible and easy to use with an intuitive interface.

There is still an underlying logic and structure to the rules, but there is no longer a class of materials concept; bibliographic entities are considered to have particular characteristics which are included in the description as required. Conventions such as chief source of information and GMD have been generalised to apply to all types of material. The primary focus in cataloguing an item is its intellectual content, not its physical manifestation.

The MARC format has been adapted to allow multi-level description: a solution has been found at last to the multiple versions problem.

Cataloguing is moving into another golden age as the demand increases for specialists to filter worthwhile resources from an increasingly complex and disordered information universe. Precision in description is seen to be the only way of achieving this.

Scenario 3: Postmodern chaos

Libraries have virtually disappeared and can only be found in remote and aged communities. People access information resources and entertainment from their homes. Print and other tangible formats are the exception.

AACR is long out of print and nobody has bothered to archive the electronic version.

The postmodern ethos has overtaken society and style is always preferred to substance. Thus the universal dictum has become "near enough is good enough." Some metadata standards exist in particular domains where precision is still important (such as medical science), but the general public - and most undergraduate students - are satisfied with anything remotely related to their topic of interest.

Anyone left in the library profession has very sensibly retrained so that they can move on to something else.

Any of these scenarios (or infinite variations on them) is possible, and I defy anyone to predict what it really will be like in ten years time. Future-gazers have a very poor record of success. In this age of rapid technological change we are lucky if we can correctly predict three years ahead. AACR will be the product of its time; given sufficient resources and the support of its profession, it will continue to meet the needs of most of its users most of the time.

It will be a challenging time for JSC. The members and the constituent bodies they represent must try to strike a delicate balance between responding to immediate needs while at the same time taking a strategic, long-term view. It must also continue to solicit a wide range of opinion from the cataloguing community but avoid the paralysis of indecision. The members must continue to be proactive while being sensitive to the representative nature of their roles.

Unlike metadata developers, those responsible for the ongoing development of AACR2 are constrained by the weight of what already exists. Consideration of any major change must take into account the impact on existing catalogues and systems. Developments that are theoretically and intellectually desirable may be too costly to implement.

Relationship between AACR2 and metadata schemes

The point has often been made that the distinction between "traditional" library cataloguing and metadata is artificial - that they are both performing the same function, but at different levels of complexity and specificity. However, although cataloguing is metadata, metadata (in the narrow sense) is not cataloguing. It does not go anywhere near meeting the functions of the catalogue as commonly understood.

It has been interesting to watch the evolution of the Dublin Core standard from one originally conceived for use by authors of electronic documents to a more formal standard for use by specialists, including librarians, in a retrospective mode - following the model of traditional cataloguing. The tug-of-war between the "minimalists" (who want to preserve its simplicity and usability) and the "structuralists" (who advocate the use of qualifiers to improve precision) seems to be edging towards the structuralist camp. As the standard develops and the number of Dublin Core records grows, its developers are starting to come up against some of the conundrums of information storage and retrieval. The natural tendency is to refine the standard, but the result may be to move it so far from the original concept that it is neither one nor the other.

This is not to say that metadata does not have a role to play in the organisation of the bibliographic universe. Obviously it is impossible to catalogue even a small proportion of all the electronic documents on the Internet. Using a form of simplified "cataloguing" is a good way of meeting an immediate need. My concern is that metadata is being promoted as the ultimate solution. I think that there are many in our profession who sincerely believe that AACR2 is obsolete and that metadata will become the new standard. This is a very attractive proposition, as metadata is relatively easy to create and does not require the expertise of professional cataloguers.

Let us imagine for a moment that these predictions prove to be correct, and that metadata standards continue to develop and move towards the level of precision that AACR2 currently achieves. Before long the same issues will surface. The same questions will be asked. Another group of people will be trying to decide when to create new records, how to deal with multiple versions of a work, how to describe resources that change over time, and so on and so on. The problems inherent in cataloguing Internet resources - their instability, their lack of boundaries, etc - are problems for metadata developers as well.

Ultimately metadata developers will have to confront the issue of authority control for names of persons and corporate bodies. A reliance on keyword searching alone will prove to be unworkable, as standardisation and consistency of access points are essential for effective searching. In library catalogues this has been achieved through the application of the rules in part II of AACR2. If we ceased to apply these rules then bibliographic chaos would result.

At a generalised level there is correspondence between AACR2 and metadata standards, and one can be converted into the other. However, when AACR2/MARC is converted into metadata there is significant loss of precision. When metadata is converted into AACR2/MARC it does not immediately become the shell of a full record. Considerable editing of the content must take place to make it conform to AACR2 standards.

Integrated approach to accessing bibliographic resources

In many libraries the provision of access to electronic resources has become the responsibility of reference and systems librarians. Library Web sites commonly contain lists of electronic resources selected for various reasons; for example, the full-text electronic journals to which a library subscribes, or electronic resources that reference librarians consider to be of interest to library patrons, usually arranged in broad subject categories. The lists vary in fullness of description; they may be simple title lists, or in some cases they may contain records that bear some resemblance to catalogue records. Links directly to the resources themselves are provided. In most cases reference librarians are responsible for the creation of the records, often acting quite independently of the technical services area of the library.

These lists are little more than parallel "catalogues." It seems odd that electronic resources should be considered so different from their tangible counterparts that the provision of description and access has moved away from the area with the expertise to provide those services, namely the cataloguing department. It also seems odd that print and other tangible resources are given full cataloguing while electronic resources are given brief and non-standard treatment. It is a kind of discrimination!

As a result users are denied integrated access to the range of resources available. Even those users who are aware of the existence of these parallel "catalogues" must search in two different places and adjust to two totally different approaches. How much more useful it would be if records for the electronic resources were routinely included in the library catalogue (with the MARC 856 field linking directly to the resources). Lists on the library Web site should be constructed from the existing cataloguing data, thus eliminating duplication of effort and ensuring consistency in description.

The way forward

Michael Gorman has identified four possible approaches to the cataloguing of Internet resources: full cataloguing using AACR2 and MARC; enriched Dublin Core records (the structuralist approach); minimal Dublin Core records: and, reliance on unstructured full text keyword searching (Gorman, 1999, p. 20). He proposes that the level of cataloguing applied should depend on the relative value of the resource. He accepts that determining the inherent value of an electronic resource will not be an easy task, but maintains that this is the only way of resolving the debate about whether to apply traditional library cataloguing standards or those for the Dublin Core.

From this perspective the nature of the debate changes completely. It is no longer a question of how to catalogue Internet resources; the rules in AACR2 are perfectly adequate, and metadata schemes provide a measure of access. It is a question of what to catalogue. The decisions to be made relate to collection development, not cataloguing. Library collection development policies should include criteria for the identification of Internet resources of continuing value so that records for them can be included in the catalogue.

In the end it is a question of resources. Cataloguing requires expertise and time, both of which are expensive. The library profession must look to the model that has served us so well - that of cooperation and sharing. Contribution of catalogue records for Internet resources to shared databases must be encouraged and rewarded.

The profession must recognise that there is no need to reinvent the wheel, for the wheel that already exists is still rolling along quite nicely. To return to the analogy of the bibliographic universe, it seems to me that AACR2 is a little like gravity. Gravity is invisible and therefore somewhat ignored - not many of us stop to think about it in the course of our daily lives - but in reality it is the force that holds the universe together. Similarly the rules in AACR2 impose structure and order; without them the bibliographic universe would degenerate into chaos.

May the force be with you!

Recommendations

1. That JSC, in conjunction with the Committee of Principals of AACR, continues to proactively pursue revisions of the code to accommodate changes in the environment.
2. That JSC, in conjunction with the Committee of Principals of AACR, continues in its quest to re-examine the underlying principles of the code to determine whether they are capable of supporting ongoing change.
3. That JSC continues to strive to expedite the rule revision process.
4. That the co-publishers of AACR2 explore the potential of the electronic version of AACR2 with a

view to making it more flexible and user-friendly.

5. That the library profession throws its full support behind the continuing development of AACR.
6. That libraries be encouraged to include in their collection development policies criteria for identification of Internet resources of continuing value with a view to giving them full cataloguing.
7. That cooperative cataloguing of electronic resources be encouraged.
8. That libraries be encouraged to provide integrated access in their catalogues to both electronic and tangible resources.

References

Delsey, Tom (1998). *The logical structure of the Anglo-American Cataloguing Rules - part I.* Available on the AACR Web site at: http://www.nlc-bnc.ca/jsc/index.htm

Delsey, Tom (1999). *The logical structure of the Anglo-American Cataloguing Rules - part II.* Available on the AACR Web site at: http://www.nlc-bnc.ca/jsc/index.htm

Gorman, Michael (1999). Metadata or cataloguing?: a false choice. *Journal of Internet Cataloging.* 2, 1, 5-22.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional requirements for bibliographic records: final report.* Munchen: Saur.

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Comments by
Lynne C. Howarth, Dean

Faculty of Information Studies,
University of Toronto, Toronto, Canada
in Response to
AACR2 and Its Place in the Digital World: Near-term Solutions and Long-term Direction
A Paper by
Ann Huthwaite, Bibliographic Services Manager
Queensland University of Technology Library, Kelvin Grove, Australia

Final version

1.0 Introduction

It is an honour to share this podium with so many distinguished colleagues and to be assigned the distinct pleasure of responding to a paper by an individual who spells "catalogue" with the same ending as I! Ann Huthwaite begins by saying that she wears many hats, though her role as Chair of the Joint Steering Committee (JSC) predominates throughout most of her discussion. I confess to having read the paper with two distinct perspectives foremost among my reactor "hats". First, as past Chair of the ALA CC:DA Task Force for the Harmonization of AACR2 and ISBD(ER), and as a member of the IFLA ISBD Review Group, I am sensitive to the *AACR2* and ISBD approaches to bibliographic control. As current Chair of the IFLA Section on Cataloguing Working Group on Metadata Schemes, and as Co-chair of the IFLA Metadata Discussion Group, I appreciate the role and value of metadata applications within a variety of unique, diverse domains.

My own academic research and teaching in the areas of bibliographic control, and metadata schemes[1], respectively, underscore my regard for the two key components addressed in Huthwaite's paper, though I sometimes view that dual interest as a kind of "fox in the henhouse" situation.

2.0 Synopsis of the Highlights of the Paper

2.1 Changes in the Bibliographic Universe

Huthwaite's paper begins with a description of the *place of AACR2* in the digital world. She underlines the many and significant changes in the "bibliographic universe" (p. 1) that have occurred since 1978 - when AACR2 was published. I have loosely categorized these changes into three areas, namely, "media formats", "end-users", and "bibliographic players and arenas".

In the more than two decades since, the predominance of print as the medium for recorded information has been challenged by a variety of new media types and information technologies. These emerging formats are, in Huthwaite's words, "notoriously difficult to pin down", inherently "unstable", and hard to "squeeze into the monographic mold" (all quotes from p.1). Electronic documents are characterized as having fuzzy or indefinite boundaries, in contrast to books or other physical media that normally have an identifiable beginning and end.

The paper describes users in the new digital world as having different expectations, best summarized as, "I want it now and in full text". While they may be more technology literate and search savvy, they are also confounded by information overload.

The third key area in which Huthwaite sees substantial change, is that of bibliographic players - enter metadata developers - and operational realities - exit public funding for significant, creative bibliographic control initiatives. She reminds us that the metadata experts that are emerging in other domains and disciplines are, " ... now debating the very issues that cataloguers have been dealing with for generations." (p. 2)

Commentary:

While I dispute neither the extent nor the significance of the post-AACR2 changes that Huthwaite relates, I suggest that what she is describing is consistent with other periods of intense code revision, whether that activity was spurred on by a proliferation of information resources, emerging media formats, new technologies, or new operational realities. Gorman's paper, *From Card Catalogues to WebPACS: Celebrating Cataloguing in the 20th Century*, provides an eloquent summary of successive - though not always progressive - code revision intent on standardizing cataloguing, and, more recently, on facilitating universal bibliographic control.

2.1 Near-term Solutions

2.2.1 Issues addressed to-date

In terms of near-term solutions for positioning *AACR2* in the digital world, Huthwaite begins by describing some perceived short-comings of the code. Briefly summarized, these include:

- the appropriateness of applying *AACR2* rule 0.24 (i.e., the provision for basing the description on the item-in-hand) to remote access electronic resources (i.e., no physical item-in-hand)
- problems with accommodating new types of media with characteristics of more than one class into one single class
- concerns that *AACR2* rules are not sufficiently flexible for "chameleon" formats or for seriality in the e-environment
- the so-called "Multiple Versions problem"
- the "glacial rate" of rule revision
- problems with the underlying structure and internal consistency

Huthwaite goes on, however, to paint a compelling picture of an agenda of changes to *AACR2* that the JSC has pursued tenaciously since the Toronto Conference in the Fall of 1997. In a mere three years, JSC has initiated a slate of rule revisions that could be characterized as being appropriately responsive to changes in the bibliographic universe, as well as more accommodating and flexible towards emerging new media formats. Such changes include (though are not limited to) the following:

- rewording of AACR2 rule 0.24 (content *and* carrier)
- an Appendix of guidelines for dealing with "multiple versions"
- revisions to Chapters 9 (aligning ISBD(ER) with AACR2), and 12 (better accommodating "seriality")
- inclusion of a statement of *AACR2* principles in an expanded introduction to the code

Efforts to tackle the perceived problems with the underlying structure and internal consistency of the code have been addressed by JSC's commissioning of a logical analysis of *AACR2* Parts I and II (Delsey 1998 and 1999 cited in Huthwaite references, p. 11), with further work being anticipated on recommendations arising from these entity-relationship modeling projects.

Commentary:

The vigorous - one might venture, unprecedented - agenda of code revision in which the JSC has been engaged over the past three years has likely rendered the adjective, "glacial", less applicable than previously. Such revisions have entailed extensive consultation across international boundaries, and active consensus-building among numerous groups - themselves comprised of individuals with sometimes differing opinions regarding the most appropriate course for rule revision. International consultation and consensus building speak to the strength of *AACR2* as a model for standards development, while also hinting of the potential pitfalls of following such an inclusive, time-intensive process.

2.2.2 Issues still to be addressed

Huthwaite speaks to JSC commitment to maintaining momentum with so many revisions planned and underway, while also noting potential problems with coordinating an increasingly complex program of work. She reminds us, rightly so, that the Joint Steering Committee is comprised of individuals juggling *AACR2* with numerous other responsibilities, but underscores her expectation that, "... *AACR2* will undergo significant change over the next five years" (p. 7), by listing a number of "intractable" problems which the JSC is committed to tackling over the near-term. These relate, in particular, to ongoing concerns with the internal structure and consistency of *AACR2*, and to problems characteristic of electronic resources, as follows:

- Structural issues
    - reorganizing Part 1 according to ISBD areas of description (Delsey) or identifying rules within Chapters 2-12 that could be generalized and moved to Chapter 1 (JSC)
    - GMDs - considering an order of precedence, compound GMDs, or no GMDs
    - End-user customizability of AARC2 and the potential of the electronic *AACR2*
- Particular problems inherent to e-resources
    - Defining boundaries
    - Variations in presentation
    - Frequency, nature, and degree of change to the same document

2.3 Long-term Direction

Stepping out of her role as JSC Chair, Huthwaite elaborates on her personal vision of three distinctly different future scenarios for *AACR2*. These I describe, simply, as (1) status quo, (2) a new "golden age" of cataloguing, and (3) let's all just go home, now! With her JSC hat firmly repositioned, Huthwaite suggests that, regardless of which direction *AACR2* ultimately evolves, the code will require strategic balancing between a number of challenges and opportunities.

2.3.1 Challenges

AACR2 - and its drafters, the JSC - must continue to juggle responsiveness to change with constituent input. While eager to embrace new media and information technologies, the code must be respectful of long-established systems (in the broadest sense), inflicting minimal disruption to operational imperatives, and carefully weighing costs to benefits. Or, in Huthwaite's words, "Developments that are theoretically and intellectually desirable may be too costly to implement" (p. 9). While those who are in the process of developing metadata standards may not yet face the same constraints, it appears that they are having to contend with many of the same enduring problems with information storage and retrieval that have confounded the bibliographic control contingent across time. For example, the challenges posed by electronic resources retrieved through remote access are as relevant to metadata developers as they are to the JSC.

2.3.2 Opportunities

There are, however, a number of strategic opportunities for *AACR2* presented by the emerging digital world. As Huthwaite notes, where institutions are offering Library Web sites of selected electronic resources in parallel with MARC-enabled catalogues, end-users are denied integrated access to the full range of tangible and intangible materials available to them. There is clearly a role for *AACR2* to play in providing the consistency in description that can greatly enhance the search and retrieval experience. Huthwaite maintains that determining the inherent value of an electronic resource (Gorman, 1999, as cited in Huthwaite references, p. 11) will help resolve whether to apply the rules of *AACR2* or to use another metadata scheme, such as Dublin Core. Thus, the question shifts from "how" to catalogue to "what to catalogue" - a collection development matter.

Commentary:

In my opinion, this scenario offers an opportunity for cataloguers to become more active partners in the acquisition and processing of a library's electronic resources. Rather than simply reacting to a request to catalogue an item, cataloguers can participate in assigning value to an electronic resource, and to determining the descriptive treatment to be accorded that object.

3.0 Further Considerations and Concluding Remarks

All in all, I think that what Huthwaite's paper highlights, and what this Bicentennial Conference explicitly demonstrates, is that we have expanded our questions around cataloguing codes from those of "what" and "how" to those questions of "where", "when", and, most importantly, "why". Having said that, and recognizing that *AACR2* is the focal point of Huthwaite's discourse, I would like to step back and pose a number of practical and philosophical questions which the paper particularly evoked for me. I would suggest these as background or further consideration to inform the eight recommendations presented by Huthwaite at the conclusion of her paper.

Question 1: First and foremost, what is the place of *AACR2* relative to other codes and standards, including metadata? At a time when there is obvious and increasing interest in standardization and consistency [2], it may be advantageous to ensure that the nature, purpose, and role of *AACR2* are clearly delineated, particularly in reference to the ever-expanding digital environment. There may be areas, not previously identified or articulated, where *AACR2* represents the most appropriate standard to apply. Understanding the inherent, relative position of the code will be an important first-step towards determining future directions for the code. A particularly useful model for how rethinking and reframing an existing standard can expand its relevance and range of application, is that of the Dewey Decimal Classification (DDC). Viewed (and positioned) most broadly as a subject access system, rather than from the more narrow and limiting perspective of a "shelving device", the DDC is being implemented in web-based knowledge repositories, portals, and intranets as a highly sophisticated, flexible information storage and retrieval tool. Extending beyond its traditional library base, and within the knowledge management arena in both public and private sector organizations, the DDC is being innovatively deployed as a universal taxonomy. One can envision similar opportunities for extending the base of applications for *AACR2* with careful, creative rethinking.

Question 2: How can the bibliographic control community ensure that it has, not only a place, but also a "voice" at the metadata developers' table? Others have noted, as Huthwaite observes (p. 9), that library cataloguing codes and metadata schemes perform the same function, but at different levels of complexity and specificity. Assuming that the two sets of standards do or can complement each other, it will be useful to have the relative merits and applications of *AACR2* represented wherever metadata standards are being developed. For example, would it not be appropriate to ensure formal JSC participation in the W3C, especially in the discussion of the Resource Description Framework?

Question 3: To what extent is the "vision" and "reach" (relevance) of *AACR2* restricted to the Anglo-American context only? The code was developed within, and reflects the values of, the cataloguing cultures of North America, Great Britain, and Australia. Gorman in his keynote paper recounted the history of bringing the "Anglo" and "American" perspectives together in the 1978 code. While respecting the roots of *AACR2*, it may be advantageous to review how the code reflects a set of cataloguing norms or values that may be restricting a broader relevance or applicability for the standards.

Question 4: How far is the *AACR2* community willing to support the "international" focus and applicability of the code? This was a key question at the Toronto Conference in 1997, but one which ultimately received a lower priority than those regarding "content versus carrier", or seriality, or the underlying structure and principles of *AACR2*. Nonetheless, as revision of the International Standard Bibliographic Description (ISBD) for various formats continues - as occurred with the publication of the

ISBD for electronic resources *(ISBD(ER)*, and subsequent JSC consideration of harmonizing *AACR2*, and *ISBD(ER)* - the degree to which *AACR2* will and *should* incorporate or reflect international cataloguing by having descriptive rules based on the ISBD will recur as a key question. Will a restrictive "Anglo-American" focus cast ISBD as Trojan horse, or is there genuine will for an *AACR2* with "global vision"?

Question 5: In continuing to revise *AACR2* within a dynamically evolving technology-enabled bibliographic environment, how best can the current adherence to principles-oriented code revision be maintained as counterbalance to pressures for systems-driven code revisions, or ad hoc rule interpretations? Changes in technical functionality may render the allure of developing a code that is more reflective, or takes greater advantage of, systems capabilities more irresistible. Throughout its history, while *AACR2* has generally responded to changes in technology, it has remained fully independent of any particular automated system. How, or how successfully this separation can continue will require some concerted thought and perhaps a resolution.

Question 6: What role should/must cataloguers play in determining/defining the "value" of Web resources? Evaluating library resources for possible acquisition has traditionally resided within the domain of Collection Development. But when cataloguers are increasingly responsible for authoring Web-enabled tools, for organizing content-intensive databases, portals, and subject-specific Websites - in some cases serving as Information Architects, or e-Content Managers - their formerly circumscribed roles are coming under review. The somewhat limited picture of cataloguers as organizers of information reacting to Collection Development selections stands in sharp contrast to the more proactive role of cataloguers as content creators working at the beginning as well as at the later stages of the information lifecycle. Rethinking the roles and responsibilities of cataloguers within the digital environment must also entail a review of their education and training needs.

Question 7: Will metadata be a "bridge" or another "wedge" between cataloguers and others? The IFLA Metadata Discussion Group is co-sponsored by the Section on Cataloguing and the Section on Information Technology. Metadata has become the focus and common talking point linking the two Sections. While the Section on Cataloguing looks at metadata from the perspectives of content description and management, the Section on Information Technology considers technical issues of metadata, such as mark-up standards, and metadata-enabled search engines and functionality. Metadata as focus provides each group with an opportunity to learn about and from one another, as well as with a forum for exploring commonalities while also respecting differences. Notwithstanding the joint programming of the IFLA Metadata Discussion Group, there remains, in some circles, a sense of cataloguing codes and metadata schemes as "two solitudes" with unique, domain-specific, and mutually-exclusive application. Is there an argument to be made for maintaining such a separation?

As a participant commented earlier, it may not be necessary to throw out the baby with the bath water, but rather, to refocus our attention on the bathtub! An apt metaphor for metadata as container. While there is no denying that some domains require unique and highly-specific metadata, as is the case with the digital geospacial domain, for example, there may be opportunities for exploring complementary applications which will mutually engage metadata schemes and cataloguing codes, as exemplified by *AACR2*. At the very least, both "camps" stand to benefit from the experiences of the other. Those in the *AACR2* community need to remain open to the potential "value add" to be derived from other metadata schemes while also acknowledging their domain-specific applications and constraints. For example, while *AACR2* is in no way sufficient to the task of describing electronic texts to the degree provided for through the TEI metadata set, the former can provide an important and concise link to the electronic publication through a library catalogue, database, or Web-enabled gateway. Used in tandem, *AACR2* and TEI offer humanities scholars targeted and effective access to important electronic sources.

For metadata scheme developers, and those who apply and use the schemes, there may be valuable lessons to be derived from the experiences of the *AACR2* community. The long-standing history of cataloguing code development and continuous revision represents decades of consensus-building, international cooperation, and vigorous advocacy on behalf of consistency and standardization. This may be a process worthy of emulation by metadata communities seeking common agreement on global interpretation and application of general and domain-specific schemes. Whether the paths of *AACR2* and

BEST COPY AVAILABLE

metadata schemes remain parallel but divergent, or convergent and complementary must ultimately be determined with the best interests of a diverse and globally-situated population of information seekers firmly in mind.

1.) The author thanks the Social Sciences and Humanities Research Council of Canada (SSHRCC GRG # 410-99-1287) for its generous assistance in funding the metadata mapping and modeling project currently in progress.
2.) Witness the demand for expertise in developing and maintaining taxonomies for enterprise portals, or for standardizing vocabulary usage in intranet knowledge repositories.

Library of Congress
December 19, 2000
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Sally McCallum

Chief, Network Development and MARC Standards Office
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4160

Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives

About the presenter:

Sally McCallum is presently Chief of the Network Development and MARC Standards Office at the Library of Congress, the Office responsible for the maintenance of the MARC21 formats and a number of other interoperability-related standards such as an XML version of MARC, the Z39.50 Information Retrieval protocol, the Encoded Archival Description DTD, and the HTML standards used internally by LC for its web site. She has been an active participant in many organizations and working groups over her more than 20 years at LC, including the MARBI Committee of the American Library Association; boards and committees of the National Information Standards Organization (NISO); committees of the International Organization for Standardization (ISO) that develop standards for libraries and information services; and the Program for Cooperative Cataloging (PCC) She has also been very active in the International Federation of Library Associations and Institutions (IFLA), chairing the Professional Board and the Standing Committee on Information Technology and serving on format related committees responsible for the UNIMARC format. She has published a number of articles on standards and networking. McCallum has a BA from Rice University and an MLS from the University of Chicago.

Full text of paper is available

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

How will MARC accommodate changes to AACR2 and developments in alternative bibliographic control tools (DC, XML, RDF)? With the recent publication of *MARC 21*, MARC enters the new millennium as a proven and robust standard with a rich history of application in library OPACS and WebPACS worldwide. MARC was developed 30 years ago, long enough for the usefulness of a common format for data exchange to be appreciated and capitalized upon. Its broadly participatory maintenance process, well supported maintenance, and stability have enabled libraries to drastically cut cataloging costs AND to vastly enhance retrieval tools through automation of the catalog. But interoperability made possible by the format is ultimately dependent on the "interoperability" or compatibility of the data it carries. The cataloging conventions can make or break these savings and advances, and can be more critical than the actual carrier format.

This paper deconstructs the "MARC format" and similar newer tools like DC, XML, and RDF, separating structural issues from content-driven issues. Against that it examines the pressures from new types of digital resources, the responses to these pressures in format and content terms, and the transformations that may take place. The conflicting desires coming from users and librarians, the plethora of solutions to problems that constantly appear (some of which just might work), and the traditional access expectations are considered.

Paul Weiss,
commentator
Manager, Conversion and Database Services
Innovative Interfaces, Inc.
5850 Shellmound Way
Emeryville, CA 94608

About the commentator:

Paul J. Weiss is currently Manager, Conversion and Database Services Unit, at Innovative Interface, Inc. Previously he was Head of the Bibliographic Control Section at the University of New Mexico, Systems Librarian at the National Library of Medicine, and Monographs and Computer Files Cataloger at Cornell University. He has been active in standards work for the past 15 years, including service on the American Library Association's Machine-Readable Bibliographic Information Committee, Subject Analysis Committee, and Committee on Cataloging: Description and Access. He currently serves as the ALA Representative to NISO. He received his B.A. in linguistics from Cornell University and his M.L.I.S. from the University of California, Berkeley.

Full text of commentary is available

Library of Congress
January 29, 2001
Comments: lcweb@loc.gov

Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives

Sally McCallum
Chief, Network Development and MARC Standards Office
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4160

Final version December 2000

Introduction (1)

The Library community has been using the MARC format as a bibliographic data exchange structure for 30 years. Its data supports simple and complex retrieval by end users of information, it is the foundation of cost-saving copy cataloging, it is the underpinning to the proliferation of interchangeable and modular bibliographic control systems that have enabled libraries to automate in an integrated manner, it is the anchor around which a rich array of tools that help libraries do their work have been built, and it has become a language that thousands of bibliographic control staff use to input and discuss control issues. This usefulness has been built over the years as systems, tools, training and globalization made the MARC standard into a keystone for automation and development. While the MARC format is simply a communications format it turned out to be the key standard and has been used in innovative ways inside and outside of systems to provide users with retrieval and services unheard of 30 years ago.

During those 30 years information resources have also gone through evolutions. In the early MARC days the challenge was achieving consistent bibliographic control of textual material, then as cataloging standards for non-textual resources were developed, the format was constantly enhanced to accommodate them -- maps, music, graphics, moving images, etc. By the late 1970s the computer file became an important library resource and various forms of expression -- text, graphic, cartographic, and sound -- began to appear in electronic and digital forms as exemplified by the digital CD that has almost completely replaced the "vinyl" phonodisc. But an explosion took place in the last 10-12 years with the development of a communications vehicle for electronic data, the Internet, and subsequent breakthroughs in systems and software that established the web environment.

Today's challenge is to provide appropriate tools for finding and retrieving the burgeoning web resources, and today's conference is looking at past practice and tools and their adaptability and applicability to the new environment.

Part 1: Extending MARC for Web Resources

By the late 1970s libraries needed to control and provide access to machine-readable data files, "MRDF", along with other non-electronic resources. As a result, through the help of an American Library Association (ALA) committee of specialists, data elements were added to the MARC format to describe these files, indicate their sources, and provide for their standard numbers. With the advent of personal computers and CDROMs in the early 1980s additional elements were added, especially to indicate physical attributes and requirements of the media. At that time the new media were considered still to be in an evolving state, so there was caution about adding too much specialized descriptive information to the bibliographic record. What would have lasting usefulness for retrieval and deployment of the media was not well understood.

Then in the early 1990s came extensive development of *online* electronic resources, including gopher technology followed by the web. There was an immediate need to sort out description issues for these resources, and especially to provide electronic links from the bibliographic record to the actual resource -- which might be local or remote. In 1993, even before the Uniform Resource Locator (URL) addressing schema was completely developed, a MARC field (Electronic Location and Access, 856) was established for information what identified the path to a resource. That field has been adjusted at least annually since then as the Internet and web environments changed and matured -- the URL became an Internet Engineering Task Force (IETF) recommendation (a type of Internet standard); access methods expanded and had to be accommodated; file format types were better understood; and work on the Uniform Resource Number (URN) was completed. More recently, with the increasing numbers of web documents and the further development of linking to support navigation, the inclusion of URLs/URNs for related materials has meant allowing them to occur as needed in various MARC fields. This is an area where changes will continue to occur.

Spearheaded by the MARC format maintenance process, major discussions took place in the mid 1990s about the "real" nature of electronic information. Initially electronic data was treated like a new form of expression, first called "data files", then "computer files". As more material appeared in electronic form, the bibliographic community revised its views, changing terminology again to "electronic resources" and recognizing that in most cases computer files are a media that carries information in a recognized form of expression, such as textual, cartographic, graphic, etc. The MARC format was adjusted to be able to encode this view, enabling users retrieval of works across physical media, for example, a specific text in print, microfiche, and electronic form.

Recent work has identified a new "issuance" pattern for electronic resources that frequently change, coining the term "integrating" to describe them and identifying their differences from traditional serially issued and monographic material. MARC discussions for accommodating the integrating resource model in the format are being held in parallel with the bibliographic discussions so when the community decides on the final requirements the format can quickly respond.

The discussions in the library community have thus focused on description and retrieval of electronic resources *along with* other physical information media, rather than separately. Distinguishing characteristics of electronically presented material are identified so that established cataloging principles for these other media can take electronic documents into account.

It is clear that the MARC format can provide a vehicle for the description of web and networked resources, and has kept up-to-date with developments in the medium. This is very positive and reassuring given the large community investment in MARC-based control -- the vast body of important non-electronic resources to which MARC is the key to interchange and the cost saving bibliographic services and tools that have been built on the standard format. But there are two factors that point to possible new directions. The first is the enormous and growing number of electronic resources and the impossibility of applying all of the current cataloging practices to them, and the second is the potential to unite electronic resources with cataloging data in new ways to assist retrieval. These make it important that the bibliographic community experiment with three key aspects of its bibliographic control environment:

- differentiation and selection of resources for levels of control,
- reevaluation of descriptive content requirements for cataloging, and
- the exchange record format structure.

Part 2: Alternatives for Control of Web Resources

Unbundling "MARC"

MARC is a generic name used for a bundle of components that come together to create MARC cataloging (or "metadata") records. These components are: structure, content and markup. The MARC format employees a standard structure to form a container for the cataloging content, which is controlled by a number of content standards. The markup (or tagging or content designation) is designed to be data identification oriented, often indicating the semantic relationships of elements of the cataloging content.

MARC structure. The underlying concrete syntax used for the MARC record is a simple "introduction - index - data" structure specified in the ANSI/NISO X39.2 and ISO 2709 standards (2). The standards dictate the length and some of the content of the introduction (the MARC Leader) and a few rules and options for construction of the index (the MARC Directory) and data fields. The MARC format implementation of ISO 2709 specifies a few additional rules about exactly how the index entries are to be constructed for the MARC format (tag length, starting character position, etc.) and how data fields are configured (positional fixed, subfielded variable length, indicators, etc.).

When the format was developed the driving forces were: to efficiently accommodate variable length data, to enable easy selection of subsets of data elements, and to provide sufficient semantics (parsing and markup) to support data element identification that would open up many possibilities for retrieval, internal system record configurations, and data manipulation. The MARC format also needed to be able to accommodate various bibliographic data models. In the library context this meant data constructed according to various national or earlier cataloging rules, in addition to new bibliographic models in the future.

The MARC record structure has been constant for 30 years. This (along with stability of the tagging and the content rules) has been a strong factor in the proliferation of systems and services related to bibliographic operations and the enormous interchange of records currently occurring among systems. This structure is, however, interchangeable as is illustrated by the fact that many (most?) systems do not hold records in the MARC format structure but treat it as a communications format, the intended use. Even the "MARC displays" and MARC-centric input templates so common in automated systems are not actully MARC structure but a layout of the markup components of the record.

MARC record content and markup. The goal of the MARC record content is broad -- to describe many facets of a resource in support of multiple purposes. The content is formed by this multiplicity of uses, the descriptive standards, and the perceived needs for consistent retrieval. The possible content has grown over time as the functionality that the exchange record was expected to support grew. The content has also been driven by changes and differences in content rules, causing new elements to be defined to identify new ways of expressing information. The overriding purpose of the record has been general resource discovery, although precise identification, selection, acquisition, item control, and preservation are among other basic functions the record supports. The uses and relationships behind the data in the bibliographic record have recently been analyzed in the study *Functional Requirements for Bibliographic Records* (3), sponsored by the International Federation of Library Associations and Institutions (IFLA). These functional requirements, which are already a major factor in metadata work, are no doubt being described and analyzed by other papers at this conference.

The MARC record data content, is largely driven by external standards that have been collaboratively developed and widely adopted by the bibliographic community over many years -- the International Standard Bibliographic Descriptions (ISBDs), Anglo-American Cataloging Rules, 2nd edition (AACR2), Library of Congress Subject Headings (LCSH) and other subject thesauri, Library of Congress Classification (LCC), Dewey Decimal Classification (DDC), various name authority files, various official standards (e.g., ISBN, ISSN), and requirements for cooperative projects. The markup in the MARC format that identifies the data content and its relationships is thus largely determined by these external standards in conjunction with judgement on the amount of parsing and identification needed for a machine to perform the functions that users require. MARC content and markup can be very thin or very fat, but it is always under the control of the external content rules that is tries to support.

Structure and content and markup have been differentiated above because in looking at new or different ways to support retrieval of networked resources these components need to be considered separately. Their suitability for web resources, and the impact of change, and pathways for change have different possibilities.

The information universe has never monolithically used MARC format-based exchange records, fundamental a part as they may have played. The vast and important journal literature has been generally (well) controlled by highly automated special subject domain abstracting and indexing services. Archival materials have traditionally been described in hierarchical lists, called finding aids, that are separately constructed for each archival collection. The finding aid was recently "automated" with the development of the Encoded Archival Description (EAD) DTD, for SGML or XML encoding of these aids. However when considering web and networked resources, the subjects of this conference, the terms that come up most frequently as the keys to networked and web resource discovery are Dublin Core, XML, and more recently, RDF. The first is a data element set and the latter concern syntax and semantics.

Dublin Core Data Element Set

The Dublin Core is a short name for a collection of 15 data elements that have been identified as useful for identification of networked resources. Work on the Dublin Core was initiated at a conference at OCLC in Dublin, Ohio, in 1995, with a broad group of participants from the computer and library communities. These data elements were refined and finalized at follow-on conferences and via electronic participation.

The original goal was for a simple set of elements that, if included in headers to web documents, would increase the efficacy of web resource discovery tools such as the "web crawlers", and also serve as a basis for fuller description of the resources, as might be needed if a description were to be added to a library catalog or other special metadata listing. As with any standard, propagation was difficult and the inability to have the set widely adopted for the original purpose meant that use was redirected. As a result interesting experiments have been conducted that take detailed cataloging from multiple repositories and extract the Dublin Core subset of elements from them. These Dublin Core subsets are then merged, providing top level resource discovery across repositories.

The 15 data elements were specified with a minimal stipulation of content rules, in keeping with the original intent for simplicity and flexibility of use. But with use came the inevitable push to add new data elements and qualifiers for existing ones, entity relationship information, content rules, and a markup for the 15 basic data elements. This is not surprising to the bibliographic community where there is constant pressure to extend a data element set, such as MARC, to serve additional media, functions, and new user groups, all with special requirements in addition to the core needs. Through multiple annual meetings and email discussion, sets of qualifiers and additional content rules have recently been established for the original Dublin Core. They are to be used when finer refinement of the 15 data elements are needed. The reality is that use of Dublin Core up until now has usually required the establishment of locally defined qualifiers. The agreed-upon extensions should fulfill some of those needs, but if users continue to have requirements for more detail, it is recognized that local elements will be established and used.

The use of Dublin Core data elements in the OCLC Cooperative Online Resource Catalog (CORC) project, which tries to maintain an ability to convert records between the Dublin Core element set and content rules (or lack of them) and the MARC content and rules, has been challenging. The differences in the content necessitated the extensive use of qualifiers with the Dublin core information in order to support retrieval compatible with full MARC content data. Some of the more interesting aspects of the CORC project are the special tools that have been developed to assist in automatically deriving cataloging data from the electronic resources themselves, and automatic checking of subject, classification and name authorities. These tools are not really related to either MARC or Dublin Core, however, but to the content standards and requirements of the bibliographic community.

"Dublin Core" thus refers to several things. (1) A basic set of 15 data elements for resource description with minimal content rules. The data elements are obvious enough that an author of a web document could often supply them without training. They are also common enough that they are a subset of data elements used in a variety of files and data bases, not just MARC, and can therefore be used for constructing meta meta files for first stop retrieval. (2) Dublin Core is also an officially expanded set of elements. The expansion is for qualifiers that refine the 15 basic elements and others that allow naming of the content rule used for the data. This form still does not mandate specific content rules. (3) "Dublin Core" (in quote marks) is also used to reference an input interface developed by OCLC where OCLC users can catalog resources (electronic or non-electronic). The input is via a special labeled template, called the Dublin Core template. The system attempts to impose a specific set of qualifiers and content rules to make the data compatible with data commonly found in a MARC record -- content rules that relate to AACR2, LSCH, DDC, etc., and various code lists. This system has a parallel input using MARC tagging. This appears to make the data as much MARC Core as Dublin Core.

Contributions of Dublin Core:

- widely recognized basic data element set, obvious and general enough that authors (or machines) can possibly supply them.
- through CORC and other projects, research on tools to automatically create Dublin Core data elements from electronic documents

Issues with Dublin Core:

- for the library community, insufficient consistency of data content, partly due to lack of content rules
- where content standards are specified or recommended, sometimes different from those commonly used in the library community.
- so basic that most applications need to define additional elements or subelements.

XML Structure

While Dublin Core is a set of data elements, XML (eXtensible Markup Language) is a data structure, comparable to the ISO 2709 data structure used by MARC. XML is actually a sub-structure possible under the more general data structure standard SGML (Standard Generalized Markup Language) which has a header followed by a simple repetitive tag-data form. SGML is specified in the ISO standard 8879

(4) and XML for web documents is a World Wide Web Consortia (W3C) recommendation. SGML was developed with the markup of text as the target, but has also proven useful as a programming tool. SGML has been used extensively in the publishing industry for textual material where generally corporations develop their own tag set under the structure, making interoperability impossible without first understanding the meaning of the tags.

SGML/XML tag sets. In the SGML environment a tag set with application rules is a Data Type Definition (DTD) (comparable to the MARC concept of format). ISO and other groups have tried to establish tag sets for the SGML structure that could be broadly used (similar to the establishment of MARC 21 as a broadly usable tag set for the ISO 2709 structure). Commonly used tag sets would allow easy interchange of marked up data and interpretation of the data without special intervention. ISO 12083 is one standard tag set targeted for relatively straightforward modern publications. The Library of Congress actually uses that tag set for the SGML markup of the MARC 21 documentation. ISO 12083 is widely used by publishers, but with a great deal of publishers-specific augmentation of the tagging specified in the standard. Another well publicized SGML markup is that of the Text Encoding Initiative (TEI). The TEI DTD was developed to make possible very comprehensive markup of textual documents -- markup that could support textual analysis of the documents. TEI has also met with some success, being used in many specialized text projects such as the Making of America projects sponsored by the National Science Foundaton. The TEI also influenced the DTD used by the Library of Congress for its American Memory digital projects.

But very importantly, HTML is a DTD that uses the SGML structure. It is a standardized tag set and is familiar to all as the markup predominant in the web environment. HTML has been enormously useful for documents to be displayed by web browsers because the tagging is display-oriented, focusing on the presentation aspects of a document, thus supporting display without construction of special style sheets.

The wide variation in the SGML tag sets developed, the complications of developing a complex DTD for each application, the success of HTML and desire to enhance it without going to more complex SGML structures -- along with the desire for SOME variability -- led to the development of XML as a SGML subset with special rules. XML does not require a formal DTD, just a scaled down "schema", or else a promise to be well formed. A document markup with a tag set defined for use in an XML structure should, for web purposes, be accompanied by a style sheet which will define its display to a browser. The style sheet enables the tagging to move away from the HTML presentation tagging to element identification tagging. XML does not itself specify a tag set that can be applied but a data structure open for definition of tagging that identify a given set of data elements, from general (e.g., Dublin Core) to detailed (e.g., full MARC data content).

The ISO 12083 and the TEI DTDs have now been specified in XML versions. Two other recent developments are using XML for tagging metadata that describes documents. One is the ONIX, a joint European and American publisher format for communicating book industry product information in electronic form. Products may be electronic books or printed material. Besides descriptive information, it contains data for the book selling function. The descriptive data could be a future source of cataloging data. The second development is the Open eBook initiative sponsored by the National Institute of

Standards and Technology (NIST). The Open eBook format specifies tagging for electronic book content, using HTML and XML tagging. The document description information specified for inclusion in the document are the simple Dublin Core data elements.

(Expected) positive contributions for XML

- structure very similar to that currently used (HTML) on the web -- XML has been endorsed for future use on the web.
- likely to be the structure for markup of many networked resources
- easy to establish a tag set, especially if DTD and schema concepts are not necessary for an application

Issues with XML

- if tag sets defined for use with XML structures are all different, interoperability is affected
- standards for schema and style sheets are still under development
- web use and widespread deployment still experimental and supporting tools still being developed (but at a rapid rate)

RDF

A new development, that is in its infancy still, is the Resource Description Framework (RDF). RDF is being developed by the W3C with the goal of making it a basic building block of the "semantic web", a manifestation of the web environment where the data is sufficiently related and marked up to support dynamically defining and exploiting new relationships. RDF holds a great deal of promise, perhaps some of it unattainable, but is certainly a path worth research. It provides a structured way to analyze relationships. RDF is not a concrete structure, but would logically use XML for document markup (it is itself being defined using an XML syntax) and would probably be open to externally defined content rules. It is, however, not ready for practical use but is currently an important research and development effort that may add understanding to resource description and become an important component in the future development of the Internet.

Part 3: Explorations

Web Objects and the Level of Control

Studies are just beginning to be produced that analyze the types of resources found on the web, but speaking in general terms, much of the open access web contains material that would not be collected in a library for research purposes. A generous estimate might be that 5% of the resources available on the web are of permanent research value and should especially be saved, cataloged, and preserved. (This is referred to below as the "research web material") The large part remaining is largely business information, often with a marketing orientation. (This is referred to below, for convenience, as the "ephemeral web material".)

The "Ephemeral" Web

Considering the ephemeral first, there are two basic concerns for this conference, current and future search and discovery. A presumption is made here that the current web is accessible and that past snapshots, or something comparable, of the web content are taken and stored for access in archives. For this body of material, simple resource descriptions are needed, and these descriptions are only feasible, given the vast number of documents in this class, if the resource creator takes some responsibility. This was the need recognized at the outset of the Dublin Core data element development effort -- to have a universally recognized simple set of data elements that authors capture in headers to their documents.

A simple "ideal" set of header elements with metadata about the document it sits in is needed. The elements need to be standardized as much as possible without layering too many form and content rules. Assurance is needed that the data elements and their tagging are carried over to newer markups (XHTML, XML, etc.) used for web documents.

A major question with expecting the author to include metadata is: Will the author take the time to supply it? This cannot be assured but fuller headers in only 50% of current web material would be a substantial improvement. There are numerous approaches to encourage authors to add the data. For example, an editor tool that the author can use to have the header automatically generated -- as best it can -- from the document content, and which the author can then correct. Encouragement by the library profession to major web sites to include a standard set of metadata as a requirement. Web indexers (crawlers) joining with librarians to promote awareness of the need for metadata and the benefits to both authors and users -- generally keeping the need and benefit alive and before those who can influence author behavior.

A major objection often voiced about author supplied descriptions for web documents is the tendency of some resource creators to engage in deceptive packaging -- supplying descriptive terms that will be popularly sought but do not apply to the resource. This can never be fully controlled, but a variety of efforts can mitigate it. Tools can compare content to author-supplied descriptors when web indexers skim from the metadata.

The Dublin Core set of elements are an obvious starting point for the endorsable set of basic elements. Appendix A compares the very basic elements commonly used today in HTML documents (Column 1) -- the metadata "hoped for" by popular web indexers -- with the rich Dublin Core set (Column 2). But it is also important to engage and obtain the concurrence of a wide spectrum of librarians, especially reference librarians. The many Dublin Core implementation experiments could provide data on how well the set works in retrieval. Also the MARC *content* needs to be a consideration when determining this set of descriptors. While use of the library community's content rules such as AACR2 would not be feasible for authors, content compatibility should be a maximized as far as possible as it will facilitate the variety of configurations in which this author-supplied cataloging may be useful. These will range from databases with only metadata harvested from electronic documents to catalogs that incorporate metadata related to selected web resources with non-web resources. The chart in Appendix A also gives a comparison (columns 3-5), using the simple Dublin Core as the basic match set, of the simple metadata that is currently specified in MARC and two other widely used standard DTDs (TEI header and ISO 12083). MARC, TEI, and 12083 all contain markup for considerably more metadata than Dublin Core, but have reasonable overlap with the Dublin set.

The "Research" Web

The smaller proportion of web documents that are of primary importance for current and future research will generally benefit from richer metadata that supports more precise searching, since integration of those records into the catalogs of libraries is important. Libraries will be taking steps to assure access to these resources and provide for their preservation, and they will want to continue to offer catalogs that assist the user in finding all resources, irrespective of media. Here the author-supplied metadata would be useful as a starting point for cataloging following established content rules and containing more detail.

Reevaluation of Descriptive Content

If libraries continue their experiments to save snapshots of the web (providing retrieval through document-carried metadata) while focusing formal cataloging and preservation on the part of the web judged to be of lasting research value, are there changes still to be considered for the cataloging descriptions for these resources? There are complexities in the current content of the bibliographic record for which the time may be appropriate to consider whether they are necessary in today's environment. Experts at this conference are no doubt analyzing and providing recommendations concerning many important content issues related to the cataloging of web resources, so the following relates to a special content issue that affects any cataloging *format or DTD*: the large number of data elements that are considered necessary by librarians, thus are currently supported by MARC tagging.

Intentional duplication. The bibliographic record carries many data elements in duplicate. This is largely driven by the tradition of providing information both in transcription form (as it appears on the piece) and in a normalized form. There are many examples of this in the format, for example the transcribed author name as it appears on the piece and the inverted and normalized author name, and the transcribed place of publication and the coded place of publication. Another building block of the cataloging tradition is communicating descriptive information through natural language notes for easy display to and understanding by the user. Such information has been used for retrieval but to assure consistent retrieval the information is often also in the record in a controlled or coded form. Examples are the language note and language code, and notes that identify names associated with a work and the corresponding fields with normalized forms of those names.

This duplication is defended on bibliographic grounds. Transcription is an aid to the end user to precisely identify whether the item is the one sought and to librarians and their machines to help identify automatically duplicate resources and duplicate records. Notes are end user friendly and clarify the characteristics of an item in human-readable form, while the normalized and coded data assists retrieval, especially retrieval from large files. Coded sata generally transcends language differences and can be very important for "weeding" a retrieval set through search qualification.

Multiplicity and granularity of data elements. In addition to core fields, each form of expression (text, cartographic, music, etc.) has special characteristics that can be differentiated and identified. Over time a very large number of data elements have been defined for bibliographic records for recording these characteristics. Often structured elements with each subpart individually identified, instead of unstructured notes, have been adopted because of the possibilities for retrieval precision. When content rules for descriptive data specify data elements that are made of identifiable parts, these elements are often parsed and each part is identified, even though the need for retrieval may be questionable.

The MARC tagging has expanded to support identification of duplicate and granular data elements. Although through special tagging conventions, the MARC format creates dual purpose fields and avoids some duplication, most duplication is easier dealt with if simply tagged as specified. When adding tagging for elements to MARC, tests are carried out to evaluate the need for separately identified data elements (needed for indexing/retrieval? for special display?), but the many special interests served by library cataloging data often successfully justify individual identification.

What needs to be considered -- in the context of other recommendations presented at this conference -- is whether the characteristics of the electronic material are different so that some duplication is unnecessary? Do "title pages" or their analogs in electronic documents have enough stability to make transcription as useful as it is for print or object oriented publications? Are special normalized forms of some data still as critical or is research producing information identification and searching tools that require less rigor since the whole document content may theoretically be searched? Are display, retrieval, and sorting requirements different for web resources, indicating less need for specificity?

These descriptive issues are perhaps the most difficult to address, given the large number of purposes bibliographic records are constructed to support. Even if the bulk of web resources are controlled with a simple set of data elements, with the expectation of less precise but adequate retrieval, the numbers of resources receiving detailed cataloging is large. Rather than fitting the electronic resources into the existing mold, this is an opportunity to check and confirm or change some of our approaches to description for this and perhaps other types of material.

Exchange Record Structure

The third aspect of the current environment that needs to be addressed is the exchange record structure. As indicated, MARC record content can be separated from the MARC record structure, allowing the use of different structures for exchanging the same data. This is not often considered since the products and services that are based on the MARC exchange record have developed *because of* the relative stability and predictability of the actual exchange format structure (in addition to the content). With every decade (or less) preferred data structures, possible data structures, and fashionable data structures for electronic data have changed with the development of different internal computer system architectures, so it is a tribute to the profession that automation *and* exchange have been nurtured by separating communications from internal data structures and stabilize the former. The applications can take advantage of current trends without interrupting record exchange. One good reason why the community might want to consider an alternative structure now is the apparent convergence of markup standards for the electronic document/web/Internet environment that may stabilize with XML. Cataloging data embedded in document headers or cataloging data exchanged for display through simple web browsers could be more efficiently used if transmitted in an XML structure with standard tagging and content.

In 1995 the Library of Congress, recognizing that the advent of full document markup had interesting potential for coordination with cataloging markup, gathered a small group of experts with experience with MARC, SGML, MARC in SGML, and electronic text markup. That group looked as a variety of aspects of the MARC format and made recommendations on how to treat them under the syntax rules of SGML. Out of that collaboration, the Library of Congress, with support from a contractor with special SGML expertise, produced SGML DTDs for the MARC record content and conversion programs that converts between the SGML/ISO 8879 and the MARC/ISO 2709 structures.

Since 1996, the Library of Congress has made available from the MARC 21 web site: an overview of the DTD requirements specified by the above group and two DTDs -- the Bibliographic DTD, incorporating the bibliographic, holdings, and community information format data; and the Authority DTD, incorporating the authorities and classification format data. Since early 1998, PERL script conversion utilities that convert both ways between the 2709 and 8879 structures have been freely available from the site, along with other tools for experimenting with the DTDs The DTDs have recently been specified also as XML DTDs and these DTDs will be available through the web site. The Library plans to keep these DTDs and tools up-to-date and in step with the markup standards, unless those standards become too volatile.

The experts group recommended that structural transformations be possible without loss of data. Thus, one characteristic of these MARC DTDs is that the XML tagging is MARC-like -- the tags are the same tags used within the MARC structure, with a little elaboration. For example, the tag for a title in MARC is "245" and in the XML MARC is "mrcb245". This tagging similarity is also the key to the simple structure conversion utilities. The following shows very brief MARC (Example 1) and XML (Example 2) record fragments for comparison.

Example 1 - MARC/2709

```
        [245] (part of directory entry)
        10$aData on the web:$bfrom relations to semistructures data and XML /$cSerge Abiteboul
```

Example 2 - MARC/XML

```
        <mrcb245 i1="1" i2="0"/><mrcb245-a>Data on the web:</mrcb245-a>
        <mrcb245-b>from relations to semistructures data and XML /</mrcb245-b>
        <mrcb245-c>Serge Abiteboul </mrcb245-c>
```

One of the attractions of using XML is its possible use as an input and storage structure, in addition to a communications structure. While many librarians know the MARC tags as a shorthand for data element names, there may be applications where staff do not. For example, within the Library of Congress, MARC templates with full word tagging are used in special applications for creating basic MARC records. With XML, after moving into the structure it is not difficult to convert among tag sets, especially if tag equivalencies are provided. Thus, the "mrcb245" could be converted to "m.title" if that were useful. That is another piece of an experimental tool set that the Library plans to make available.

Since XML has become popular, other XML versions of the format have been created as part of different projects indicating experimentation is taking place. With MARC-in-XML tools available, records for the "research" part of networked resources, for which full MARC content cataloging is warranted, can be either produced in XML, depending on the system, or easily converted from a MARC system to XML MARC for attachment to the XML document. This will provide a smooth pathway to what may be an eventual transition. If the XML data structure seems to have staying power -- and that is a real question given the nature and pace of change in web development -- with these tools the bibliographic community *will not have a revolution in its resource investment but an evolution*. This is important for an industry without surplus funds and with the need to keep its primary funding directed toward obtaining information resources themselves, not the conversion of catalogs.

Conclusion

This paper has discussed three avenues of exploration related to bibliographic records that the web environment invites -- sorting out the level of control for web material, reevaluating aspects of descriptive content requirements for these materials, and experimenting with new format structures. These explorations will take place with or without the participation -- or leadership -- of librarians, but they should not. Librarians need to have prominent roles in all explorations so that their cumulated knowledge and understanding of document control and discovery are built upon, not slowly rediscovered and reinvented.

As information specialists, librarians need to enhance their technical skills and collaboration skills, so they can work successfully with computer professionals, who will ultimately write the systems. As librarians they need to affirm the value of integrated access to research material -- electronic and non-electronic, different forms of expression, old material and new, etc. As responsible information servers they need to keep up with the directions technology is headed -- Will the web last? Will XML be superseded in a few years? Will there be constant costly change? What are retrieval innovations that influence record content? They need something like the following agenda.

* Apply their well honed resource selection skills to web resources, establishing general and feasible guidelines.
* For the mass of web resources, use simple descriptions and use "commercial" finding systems, but:
    Advocate for simple document descriptions embedded in web resources.
    Evaluate simple Dublin Core for that role, and submit for any well-justified changes.
    Assist in development of helpful tools to improve such simple descriptions.
* For identified research material, use MARC content -- heavy to light as needed -- but:
    Evaluate for possible unnecessary data elements complexities.
    Experiment with structural transformations such as XML for the MARC content.
    Assure that tools are readily available for conversion among structures.
*Keep the library community's understanding of the value of and commitment to standards by continuing

to work together on any changes to conventions and standards.

Footnotes

1. There are a large number of terms being used in the broader information community that often mean approximately the same thing, but relate concepts to the different backgrounds of the players. For example librarians are sometimes confused that metadata is something new and a replacement for either cataloging or MARC. Metadata is cataloging and not MARC. In this article terms based on library specialist terminology are used, with occasional use of alternative terms indicated below, depending on context. No difference in meaning is intended by the use of alternative terminology . The descriptions of the terms are indicative, not strict.
   cataloging data or cataloging content = metadata
      - used broadly, in this context, for all data (descriptive, administrative, and structural) that relates to the resources being described.
   content rules
      - rules for formulation of the data including controlled lists and codes.
   data elements
      - the individual identifiable pieces of cataloging data (e.g., name, title, subtitle) and including elements that are often called attributes or qualifiers (since generally this paper does not need to isolate data elements in to subtypes).
   relationships
      - the semantics that relate data elements, e.g., name is author of title, title has subtitle.
   content rules
      - the rules for formulating data element content
   structure = syntax
      - the physical arrangement of parts of an entity
   record
      - the bundle of information that describes a resource
   format = DTD
      - a defined specification of structure and markup
   markup = tag set = content designation
      - a system of symbols used to identify in some way the following data.
2. ANSI/NISO Z39.2, *Record Interchange Format*, and ISO 2709, *Format for Data Interchange.* The two standards are essentially identical in specification. ANSI/NISO has a few provisions where the ISO standard is not specific, but there is no conflict between the two standards.
3. *Functional Requirements for Bibliographic Records.* IFLA Study Group on the Functional Requirements for the Bibliographic Record. Munich, Saur, 1998.
4. ISO 8879, *Standardized General Markup Language (SGML).*

Appendix A - Basic Resource Description Metadata

Common HTML Header metadata	Dublin Core element	MARC core element	TEI header element	ISO 12083 element
	Identifier	Electronic Resource Identifier (856 $u)		
	Format	Electronic Resource Identifier (856 $q)	<extent>	
<title>	**Title**	Title (245 00a)	<title>	<title>
<meta name = "author">	**Creator**	Added Entry (720 $a)	<author>	<author>
	Contributor	Added entry (720 $a)	<name>	<author>
	Publisher	Publisher (260 $b)	<publisher>	<pub>
	Date	Date of publication (260 $c)		<date>
<meta name = "keywords">	**Subject**	Uncontrolled subject (653 $a)	<keywords>	<keyword>
<meta name = "description">	**Description**	Summary, etc. note (520 $a)		<abstract>
	Language	Language note (546 $a)	<language>	
	Type	Genre (655 7$a)		
	Coverage	General note (500 $a)		
	Source	Linking entry (786 0 $n)		
<ahref>	**Relation**	Linking entry (787 0 $n)		

Comments by Paul Weiss

Final version

As Sally has clearly pointed out, MARC as it is currently constructed can be used to share bibliographic information (aka metadata) about networked resources. It has been used for that purpose for a few years now. Some content designation has been added specifically for use with networked resources, most notably the 856 field. Additional adjustments can be made that will make it even more useful as we move forward. Some of these are relatively minor, adding a field here or a subfield there, and some are more major, such as embedding MARC structure in XML. However, through all these changes what remains constant is the knowledge and experience that we have gained over the years as to what is important in sharing creating, sharing, and using metadata. This is our true intellectual capital, which, I believe, is even more valuable than the actual data in our millions of records.

I believe that one of the most important points in Sally's paper and presentation is that there are multiple aspects that make up MARC, which she identifies as content, structure, and markup. Many have viewed MARC as simply markup, but Sally shows that MARC is in fact far richer than that. Indeed many other information organization tools that librarianship has developed over the years-AACR2, LCSH, DDC, etc.-have similar multiple aspects in their makeup. Our intimate expertise with these various aspects is at the root of our intellectual capital.

So what is this intellectual capital? We can start with the fact that information resources and metadata that describe them are far more complex and unruly than most people outside of our profession can even guess at. "How complicated can it be to figure out what the title of a book is?" (Well, let me show you any number of conference proceedings, agricultural research station monographs, looseleaf services, or European Union publications.) What else have we learned in our over 30 years of experience with the MARC formats and other standards that will help people identify, search for, and use networked resources on the Web?

Standards

There are too many resources physically in our libraries and now out on the Web for any one institution to create the metadata for all of it. So one of us creates metadata for a resource and we generally share it with the rest of the library community through bibliographic utilities, or by making our catalog accessible on the Web. Since we are sharing our data, and since we want to be able to use various automated tools beyond those we develop ourselves, we have created standards. These help ensure that we can read each other's data, and that system vendors have a large enough base of prospective customers to make it worth investing resources in developing systems that manipulate that data. Interoperability and a need to not reinvent the wheel dictate that we use already existing standards when feasible; the MARC formats refer to several external standards.

It can be helpful to have metastandards. ISO 2709 and Z39.2 are metastandards that provide the general structure for interchangeable records. The Backgrounds and Principles document and the Record Structure and Character Sets sections of the Specifications document form the next level of standard down. They distill the features that are common among all the MARC formats. The specific MARC formats then take that general structure and flesh out specifics for different kinds of data. Granted, in this case the standards were developed in reverse order, but faceting out the levels of structure is still valuable. In the larger Web community, SGML and XML were developed as metastandards. There is growing realization that the Dublin Core is de facto a metastandard rather than directly a standard, as many implementations of it add structure. Especially early on in the development of standards in a particular area, the creation of a metastandard for what can be agreed to by everyone allows experimentation with specific standards for specific projects. Lessons learned in these early implementations can be incorporated into a more general standard.

Sometimes it is useful to have multiple standards for the same issue; different communities often have different needs. Library of Congress Subject Headings are used by many libraries in the US, while sizable numbers use either Sears or MeSH instead. In all cases, providing a sound subject retrieval system is the goal. Many school libraries receive MARC records on diskette, while many academic libraries use ftp to move sets of records around, there being a different standard for each exchange medium.

At the same time we have learned that standards should only be developed and followed when the benefit of adhering to the standard is more than the cost of not doing so. Sometimes there is little value in standardizing an aspect of metadata. We do not require any particular structure in the content or markup of data in a General note (500); there has been no strong need articulated for doing that. Sometimes there is a value, but the cost is too high. Such value judgements may be made as a profession, as in not providing chapter-level subject access, or locally, such as which series to analyze. There are cases in which one community may find it worth standardizing and another not. The school library community finds specific information about audience level very valuable, so the 521 field was given additional structure to accommodate their needs. Meanwhile, most academic libraries, if they record this information at all, record it rather free-text. And even when standardization would be considered valuable to a user community, it may not be considered as such to the community that would need to apply that standard. Witness the situation between librarians and publishers with regards to standardizing the title of a resource in all places on and in a resource where it appears.

Experience has shown us that changes to standards need to be treated in a controlled and explicit way. The "obsolete" concept in MARC has proven to be quite valuable. Documenting changes to the format is crucial for database managers to fully understand their data. Consensus in the profession on how and when to implement changes (AACR1 to AACR2, format integration, Wade-Giles to Pinyin, even the addition of a new source code) has kept the use of our standards standard.

Resources

We have heard many times how different networked resources are from books, etc. It is important to be able to distinguish what is truly new and different with networked resources, and therefore may need new solutions, from what has precedent in the pre-networked world, and may be amenable to existing solutions. For example, the fact that networked resources often change frequently with little explicit notice given has a parallel in looseleaf material. Some of the ways we treat looseleafs may work with networked resources. If nothing else, our experience with looseleafs has taught us that there are some aspects of these variations that are more salient in identifying a resource than others. This general idea can be helpful in discussions with nonlibrarians. On the other hand, the quantity of new, thus far undescribed resources now available to our users is of such a larger order of magnitude than any of our historical backlogs, that the issue of scalability is essentially new for us.

Metadata

We know that it is important not just to have data that describes a resource (metadata), but also data about that metadata ("metametadata" perhaps). Leader byte 17 (Encoding level) describes the fullness of the record or, to some extent, our confidence in the data. Certainly this has been a useful concept for us in the past, and would be quite valuable to know about metadata for a networked resource. Another kind of "metametadata" is data about sets of metadata. The electronic file label structure delineated in the Exchange Media section of the Specifications document allows one machine to understand what it is getting from another machine. We also use data which actually describes the relationship between two versions of metadata for the same resource. Leader byte 5 (Record status) tells a system whether this is a new record, a better record, or a record to be deleted. Extremely simple and obvious concept to us, but not necessarily to other communities.

Our experience with MARC reminds us that if data is faceted out and marked up well once, it can be utilized (displayed, processed, searched on) in multiple ways, with the underlying structure of the data often transparent to a particular user. In the acquisitions arena, for example, a library staffer may input bibliographic, order, and checkin data about a new serial on one template in her library system. The system then takes that data and organizes it into three distinct but linked records. When a patron searches for that serial in the OPAC, the system brings together various pieces of data from each of those three records to provide a meaningful display to the patron.

Authority control, as many others have pointed, is one of the most important areas of expertise that librarians can share. Although much of the authority control we use in libraries is handled by other standards, MARC has its own as well, in the several code lists for languages, organizations, sources, etc. Indeed, when to have data communicated in a coded form should be thought through.

Other aspects of metadata that we have found useful to standardize include:

- repeatability and order of data elements,
- level of granularity of data elements,
- use of punctuation,
- the difference between no attempt to supply, unknown and not applicable, what characters are allowed where and under what conditions (for example only letters and numbers as subfield codes), and
- reserving space for locally defined data elements.

Next steps

So we have all this intellectual capital to share with the larger Web community. How do we go about doing that? Getting active in W3C and other organizations is certainly important. Over the years, we have made some attempts to be heard in other communities, but usually without much success. I believe that some of this is due to the sociology and psychology of our profession. We may be proud of what we do, but we only express that well to ourselves. We need to gain enough self-confidence as a profession to be able to express the value of our expertise to others. We also need to learn their lingo. Using library-specific terminology without explanation and explicit relationship to something in the other community's world will not get us very far.

Summary

Here is a summary of the points discussed above that I think we as librarians involved in bibliographic data can bring to the table when discussing access to resources on the Web with members of other communities. Perhaps one of the most important ideas we can bring is that some of the following at first glance seem contradictory, but each has its flavor of truth. We can help achieve the appropriate balance among the implications of each of these ideas to bring about an information world optimized for success.

- We librarians have knowledge and experience to bear on the issue.
- Information resources and metadata that describe them are complex and unruly.
- Standards are valuable.
- Use existing standards when feasible.
- Metastandards can be valuable.
- Different communities may need different standards.
- Standards are not always worthwhile or enforceable.
- Changes to standards need to be communicated about both beforehand and afterwards.
- Networked resources aren't really that different from traditional resources.
- There are some important differences between networked resources and traditional resources.
- Data about metadata is valuable.
- Well-structured data can be utilized in multiple ways without that structure becoming overly apparent.
- Authority control is valuable.
- There are other aspects of metadata that may be worth standardizing.

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Carl Lagoze
Digital Library Scientist
Dept. Of Computer Science
Cornell University
Ithaca, NY 14853

Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?

About the presenter: Carl Lagoze is Digital Library Scientist in the Computer Science Department at Cornell University. In this capacity he leads a number of digital library research efforts in the Department and across the university, collaborating with the University Library and Office of Information Technology. Mr. Lagoze's research is funded through a number of NSF, DARPA, and industry grants, most notably a major grant from the multi-agency Digital Libraries Initiative Phase 2. In general, this research can be characterized as investigations into the technical and organizational issues in the development and administration of distributed digital libraries. The recent focus of this research is on policy: What are the policies that need to be asserted to ensure the reliability, security, and preservation of content and services in distributed digital libraries and what are the mechanisms for enforcing those policies? Mr. Lagoze is the co-inventor of Dienst, a widely deployed protocol and architecture for distributed document libraries. He is also the co-author of the Warwick Framework, a modular metadata model for digital content, which is a conceptual basis for the Resource Description Framework (RDF), now a WWW metadata standard. Mr. Lagoze's professional activities include serving on the advisory committee of the Dublin Core Metadata Initiative, serving on the program committee of U.S. and international digital library conferences, and numerous talks both in the U.S. and internationally on his research on metadata and digital library architecture.

Cataloging Directorate Home Page

Library of Congress Home Page

Full text of paper is available

Summary: (revised 11/1/00)

The speaker proposes that the digital context presents a dramatically new context than that which was addressed by the traditional cataloging model. Whereas the catalog has depended on relatively fixed resources delivered by a relatively stable set of role players (publishers, authors, information intermediaries), the digital context is characterized by fluidity in both content and those who provide it. The speaker proposes new roles for the catalog based on this new reality and a new data model that meets the needs of these roles. An "event-aware" model of cataloging, one that recognizes digital resources as inherently dynamic, will allow the research library to adapt to the realities of the digital millenium.

Library of Congress
November 1, 2000
Comments: lcweb@loc.gov

Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?

Carl Lagoze
lagoze@cs.cornell.edu
Department of Computer Science
Cornell University

Prepared for Bicentennial Conference on Bibliographic Control for the New Millennium, Library of Congress, November 15-17 2000

Final version

Business Unusual

Since the nineteenth century, the modern library has been the preeminent institution of responsibility and trust in the information landscape. The Catalog has done much to make this possible by providing a uniform vehicle for access and management of a variety of information resources. Rapid growth of the Internet and the revolutionary transition from physical to digital artifacts jeopardize the role of the catalog and the library institution itself. The conservative "business as usual" perspective of libraries must shift to "business unusual" – radical changes in the catalog, its role and its composition, are needed for the library to endure in the digital age.

The increasing perception of information as a commodity suggests that there are lessons to be learned from the business world. In his popular management book Clayton M. Christensen [13] describes the threats and opportunities for businesses in the face of a *disruptive* technology[1]. Whereas a *sustaining* technology improves the performance of an established product, and therefore appeals to an existing customer base, a disruptive technology brings 'to a market a very different value proposition than had been available previously'. In his book, Christensen demonstrates how disruptive technologies establish a failure framework that historically has led to the exit of established companies from a market and, in many cases, their eventual demise. .

Research libraries are unquestionably confronted with a suite of disruptive technologies, so numerous that they can be described as a *disruptive context*. The elements of this disruptive context include well-known technical advances such as low-cost computers, the availability of broadband networking in the home and office, and advances in protocols and delivery systems on the Web. In addition, there are non-technical factors such as changes in the publishing framework (e.g,, the movement to 'author self-archiving' as described by Stephan Harnad [20]), and the increasing rate of change in many fields and corresponding increasing demand for immediate availability of research results. In combination, these factors seriously undermine the practices, and in fact the *raison d'être*, on which the research library has relied for over a century.

The Catalog stands exposed to the full force of this disruption. Over the past century research libraries have expended considerable effort evolving the catalog as a sustaining technology, adapting it to new genre of materials - audio, video, and software – and new delivery systems – from cards to MARC formatted electronic records and the integrated library systems that store and provide access to them. There is no question about the high functionality of the 'cataloging product' and its success in uniformly imposing order [26] on a variety of resources to facilitate their discovery, access, and management.

Yet, the nature of disruption, as described by Christensen, is that apparent success of a product often belies fundamental threats to its viability. In the case of the catalog these threats are both intrinsic and extrinsic.

The most substantive intrinsic threat to the viability of the catalog as we know it rests largely in the costs associated with it, which by and large is a result of its complexity. While automated sharing of cataloging records has produced substantial economies of scale, the cost of an original cataloging record, for which estimates range from 50 to 110 $US [15], makes it among the most expensive tasks in the library. The increasing burden of this expense led Bill Arms [3] to question whether the current cataloging practice can continue to exist amidst relatively static library budgets and the increasing number of resources to catalog. Arms suggests that a wiser resource allocation might be the use of cheaper automated descriptive methods even though the results would be admittedly less functional.

The extrinsic threat to the survival of the catalog comes from the changing nature of information, how it is delivered, and who takes responsibility for organizing and describing it. These changes can be characterized as follows:

- *Scale* – The sheer volume of information available on the Web and the rate of growth severely stresses the economics of traditional cataloging (described above).
- *Permanence* – The impermanence of digital information defies attempts to establish fixity, which is essential to traditional cataloging.
- *Authenticity* – The breakdown of traditional publishing models on the Web disrupts conventional mechanisms for establishing the authenticity of an information resource.
- *Variety* – The rapid introduction of new genres of digital information and the demand for specialized descriptive methods for these resources undermines the notion of *uniform access* upon which the traditional cataloging model rests.

Within this changed context, various types of metadata distinct from the catalog[2] are evolving as truly disruptive technologies – often cheaper, simpler, and admittedly less functional than the traditional catalog. In contrast to the catalog record, which is a self-contained complex organizational scheme developed and maintained by a closed community of professionals, metadata in general varies across a number of dimensions:

- *specialization* – formats and schemas often reflect the needs of specific communities
- *decentralization* – production and maintenance of metadata occurs in distinct communities of expertise that rarely share common practices or standards.
- *democratization* – some metadata initiatives, notably the Dublin Core Metadata initiative, are targeted for creation and maintenance by non-professionals.

As such, metadata offers the possibility of substantially lowering the cost of describing resources and making those descriptions more appropriate for the communities that use the resources. Furthermore, a number of metadata initiatives are focusing on descriptive domains largely unexplored by traditional cataloging records; for example rights management [35].

How can the research library maintain its enduring order-making role in the face of these disruptive challenges and technologies? How must cataloging and cataloging practices change so that libraries can continue to add value to the information infrastructure? There are no simple answers to these questions. The answers, as such, must address the institutional, technical, and theoretical foundations of cataloging practice. Hopefully, conferences such as this one provide the opportunity for evaluating the challenge and developing an inventory of ideas from which the community can move forward.

My view, as presented in this paper, is that adaptation to the networked information context will require rather radical changes to the role of the catalog and the cataloging model. This view and the material presented in this paper builds on some ideas that were put forward in the recently published National Research Council study of the Library of Congress [15], in which I participated[3]. As stated in this study:

The committee understands that it will be a tremendous challenge to change the base model for metadata (e.g., from resource-centric to relationship-centric) in a world of widespread data exchanges (the MARC records that are the basis of cooperative cataloging) and reliance on turnkey software (commercial integrated library systems that are based on MARC). However, it is certain that library-type metadata practices will at some point need to be reexamined in the light of a changed world. It is certainly valid to ask when the time will come where there is sufficient understanding of this changed world to undertake such a process. It is not productive to ignore the fact that changes are inevitable and dramatic.

The premise underlying this statement is that the resource-centric descriptive model upon which current cataloging practices are built, whereby discrete descriptive records are associated with fixed information artifacts, is incompatible with networked digital information. This new context has radically different information entities, decentralized information production and management, and troublesome questions about authenticity and trust. It requires a model that can flexibly express the relationships between resources, abstract concepts, and multiple descriptions of those resources and concepts[4]. Complex relationships are not unique to the digital world – examples such as translations, editions, transcriptions, and the like – are well-established in physical genres and have bedeviled catalogers for years. The nature of networked digital information, however, greatly increases the complexity of resource relationships and demands a descriptive model that fully represents those relationships.

The goal of this paper is to examine one dimension of such a new data model – *event-awareness* – and why it must be an important component of a new cataloging model. Summarized briefly, an event-aware model raises events or state-transitions to first-class status, thereby allowing descriptive properties to be associated with these transitions, as well to the information entities that are inputs, outputs, and tools for these events. Using "translations" as an example, an event-aware model defines the translation act as a "first-class object" and associates properties such as the date of translation and the translator to that translation object.

The beginning of the paper describes why event-awareness is necessary for a new cataloging model. This necessity comes from both the nature of the digital objects that the catalog must describe and the role that libraries and the catalog need to play in the digital context. The latter portion of the paper provides the outline of an event model and how it might be used. It is not my intention in this paper to provide a complete solution to the problems facing the catalog. However, I hope that some of the ideas provided here may hint at the directions such a solution may need to take.

Why event-awareness?

What has changed in the digital milieu that makes an event-aware model relevant? This section focuses on the following issues:

- The move away from relatively fixed physical artifacts to generally fluid digital objects.
- The difficulty of establishing integrity, trust, and authenticity in the networked environment.
- The decentralization and specialization of resource description and problems of mapping amongst these descriptive vocabularies.

Fixity and Fluidity

Fixity is an underlying assumption of the traditional cataloging model. Fixity is realized in the two most significant first-class entities in the traditional model – the *work* and the *document*. The "first-classness" of these entities lies in the fact that they are the locus for association of attributes created by the cataloging process [37]. Fixing the work provides the locus for the association of time and space independent attributes such as author, title, edition, and subject. Fixing the document, as a particular space-time manifestation of a work, provides the locus for associating attributes related to publication (e.g, date) and location (e.g.. library shelf). The instantiation of a cataloging record in a library's catalog establishes another layer of fixity; the linkage between a bibliographic description, a work, and document recognized as a manifestation of that work.

What is a "document" in the digital context, how does it differ from other information objects, and what is the nature of its fixity? Michael Buckland asks many of these questions in "What is a "document"" [10]. Buckland notes how the digital world, where everything exists "as a string of bits", calls into question traditional information science distinctions between documents and other information objects (e.g., processes, images, digital artwork). If "digital documents" bear a striking resemblance to "digital museum objects" or to "digital archival objects", then certainly the descriptive distinctions between these communities need to be reconsidered[5]. David Levy writes about issues of fixity and fluidity in physical and digital documents [27]. While Levy states that both physical and digital documents have degrees of fixity and fluidity, he calls attention to the significance of "technologies of fixity". Whereas both digital and physical documents move between states of fixity, there is a marked acceleration of these state transitions in digital documents; Levy calls it "the rhythm of fixity and fluidity".

The quickening of this rhythm is sufficiently problematic to call into question the integrity of the relationship of a catalog record to a digital document, thereby weakening the base integrity of the record. Examine how such relationship between record and digital object is established in the catalog. The recommended method [33] for fixing the relationship of a MARC catalog record with a networked document is through the 856 field: "Electronic Location and Access". The predominant content of this field, given the dominance of the Web for the delivery of digital content, is a URL. The fragility of URLs, or any pointer across the network[6], is well known. This fragility may take the form of catastrophic disappearance of the referenced object (known in HTTP as a 404 error), or, even more insidious, modification of the object and resulting changes in its information content (see [30]).

A solution to this conundrum – fixing the network reference – is non-trivial. One brute force "solution" is to give up on networked references, copy the objects to a local repository, and assume responsibility for their stability. As suggested in the NRC report [15], however, an attempt to indiscriminately move the "library as container" notion from the physical to the digital world is simply not realistic. Crespo and Garcia-Molina [18] suggest another solution, using techniques such as hashing for establishing bit-wise fixity. While this may appear to be a workable solution, it fails to account for the fact that exact bit-wise correspondence is not really the issue when it comes to the integrity of the cataloging record[7]. Generally, the more important issue vis-à-vis the integrity of a descriptive record is fixity of the *meaning* of the document [30] that the record purports to describe. Furthermore, bit-wise fixity is essentially meaningless when the fundamental nature of some digital objects rests in their dynamic nature (e.g., what exactly are the fixed bits the online of the New York Times at http://www.nytimes.com).

The inherent fluidity of many digital objects suggests that a "fixation with fixity" may in fact be a red herring. My suggestion is that a more realistic approach towards cataloging digital object is to incorporate fluidity into the cataloging model itself. The record should model a digital document as a series of transition events, and should describe the nature of the events, the agents responsible for the events, and the times and places of those change events.

No doubt, this "answer" to the cataloging model opens up a number of questions that will need to be examined by the cataloging and research community:

- What is the granularity of the event record that should be recorded for digital objects? Abstractly any event can be deconstructed recursively to infinitely granular levels. The challenge in any such event model is to understand how finely granular a change history should be; the answers will inevitably be community and situation specific.
- If existing resource-centric cataloging is expensive, what are the costs of incorporating events in a new model? Like many metadata problems, there will need to be solutions that combine automated and human effort. In our Project Prism at Cornell, we are examining the use of monitoring surrogates [34] as one means of flexibly tracking status of digital objects and perhaps assisting in the maintenance of event records.

Although these and other open questions remain for an event-aware model, it does address the pervasive need to address the fluidity of a large class of digital objects. The failure of the traditional catalog to do this is a serious impediment to the transition of the library to the digital context.

A Foundation for Trust

Mechanisms for trust (and component issues of integrity, authenticity, security, and privacy), which are well-established in the bricks and mortar information context, have proven to be among the most difficult to transfer to the digital milieu. Two major national studies [16, 36] and a variety of research projects have examined issues related to how to establish trust between parties, how to be certain about the authenticity of information, how to protect privacy, how to securely protect information, and how to disseminate information in a controlled fashion in the digital realm.

What is the role of information professionals, libraries, and, in particular, the catalog (and metadata in general) in resolving such trust and integrity issues? I suggest that these organizations and tools have an essential role. Furthermore, the catalog can facilitate this role only if it has the ability to record events in the lifecycle of digital objects.

The perspectives of information professionals and researchers from a variety of communities – archival, computer science, and preservation – provide some valuable background on this issue. Picking up the theme of the previous section, the issue of fixity, or lack thereof, is a large part of the problem. As noted by David Levy [28]:

Assessments of authenticity in the world of paper and other stable, physical media rely heavily on the existence of enduring physical objects. ... What happens in the digital case if there are no stable, enduring digital objects?

Peter Hirtle [21] describes how archivists, preservationists, and librarians share the same problem of authentication of digital objects and how this demonstrates the need for a common approach. The similarity between the issues face by archivists, preservationists, librarians, and others including the museum community is a concrete example of the questions raised by Michael Buckland in "What is a Document" [10]. The issues prevalent in each community merge as their individual media are commonly represented as bits on disk or over a network.

One approach from the archival perspective, advocated by David Bearman[8] [4], is for trusted custodial agencies to maintain metadata that records the provenance of the digital object. Bearman and his partners stress the importance of custodial control over provenance metadata, in contrast to control of the objects themselves. He reaches a conclusion about centralized storage of digital objects that sounds very similar to that of the NRC Library of Congress study [15]:

Archivists cannot afford – politically, professionally, economically or culturally – to acquire [electronic] records except as a last resort.

In later work [5] Bearman describes a metadata model for such a task – one that has a strong event orientation. Paul Conway [17] reaches a similar conclusion for the preservation community, stating that the solution to establishing integrity of digital objects lies in "documenting successive modifications to a given digital record".

We see in all of these statements the common argument that unlike physical objects, where authenticity is sometimes derivable from the object itself[9], authenticity of digital objects can generally only be established by endowing the objects with metadata, which is then maintained by trusted institutions. Clifford Lynch [29] addresses this *trust* issue directly and describes how all assertions of authenticity for digital objects are grounded in levels of trust:

…there is no question of authenticity through comparison with other copies; there is only trust or lack of trust in the location and delivery processes and, perhaps, in the archival custodial chain.

Lynch points out that there are a number of existing developing technologies that assist in establishing trust, but that all of these technologies recursively reduce to institutional trust (e.g., the institution or combination of institutions from which a provenance chain was derived); in other words, trusting the institutions that hold custody over the metadata establishing provenance.

How does this all translate to the role of the library and the catalog? The rapidly growing dependence on (born-again and born) digital information through society – in schools, business, education, and the like – presents a large-scale authenticity crisis. There is a compelling need for trusted organizations to step forward with tools to alleviate this crisis. I believe an essential value-added role that the library can add to the networked information environment is to act as a leader, or at least a primus inter pares, is establishing trust. I believe that the catalog should be the mechanism that facilitates this role. To accomplish this, the catalog must be able to act as a record keeping tool; one that is useful for documenting the events that take place in the origination of and modifications to digital content.

Metadata as a cross-community activity

In the Warwick Framework [24] we advocated a modular model of metadata – individual descriptive *packages*, contributed by distinct communities of expertise, that are aggregated and associated with networked resources within a metadata *container*. This modular model is realized in the RDF (Resource Description Framework) [25], which the Web Consortium is advocating as the basis for Web metadata.

The decentralization of this descriptive model is dramatically different from that presumed by the catalog, which is generally framed as a "one-stop shopping" descriptive context under the control of a well-defined professional community. Undeniably, the centralization and well-defined control regimes of the traditional catalog generally leads to high-quality descriptive records, where quality is measured as adherence to well-defined standards and rules.

It is not productive, however, to argue platonic notions of quality in the face of two countervailing factors. First, the benefits of specialization in distributed, community-specific metadata are considerable. Although AACR2 and MARC encoding has proven adaptable for a variety of resources, it simply not capable of expressing descriptive semantics in specialized areas[10]. Any attempt to incorporate such specialized semantics in a general cataloging model would only lead to greater complexity and resulting greater cost. Second, the economics of cataloging, described earlier in this paper, make it impossible for libraries to ignore the cost savings possible by leveraging descriptive information supplied by metadata from external organizations.

What then is the distinct role of the library and the catalog in this decentralized descriptive environment? I suggest that a useful approach is to enthusiastically accept descriptive diversity and adopt a role as *mediator*. Rather than absorbing semantics (and descriptions) from distributed communities, libraries should promote the catalog as a mapping[11], or interoperability mechanism, amongst distributed descriptions. Technologies such as RDF and its schema language [8] make it possible to undertake such a mapping role amongst individual descriptions that are distributed across the Web[12].

I have no doubt that this suggestion might meet some resistance from my library colleagues whom have already been asked to accept the notion of providing access and some responsibility for content not entirely in their control. This suggestion takes the idea one step further by conceiving of the catalog as not only an access point for distributed resources, but as a distributed resource in its own right.

The existing resource-centric catalog is not an adequate basis for such semantic mediation. Scalable and extensible mapping among different metadata vocabularies will require a model that recognizes distinct entities that are common across virtually all descriptive schemas – people, places, creations, dates, and the like – and that includes events as first-class objects.

The importance of this event-awareness in the model can be explained as follows. Understanding the relationship among multiple metadata descriptions (and ultimately the vocabularies on which they are based) begins by understanding the entities (resources) they purport to describe. Understanding these entities entails a comprehension of their lifecycle and the events (and corresponding transitions and transformations) that make up this lifecycle.

This argument builds upon the following observations. Descriptive communities can be distinguished by the events that are of significance to them. For example, a community that focuses on the history of production of a film may consider the "event" associated with the insertion of a certain scene into a film significant. As a result that event may be explicit in their descriptive vocabulary -- for example, that community may have a metadata attribute that describes the date of the scene insertion. Another community, say one concerned with the presentation of that film on a screen, may consider that event irrelevant and may not be concerned with the "is part of" relationship of the scene to the movie.

A particular metadata description, a record from some community in some schema, actually refers to a *snapshot* of some entity taken in a particular state - a perceived fixity of the entity in a particular time and place that perforce elides events or lifecycle changes that are outside the domain of interest by the particular descriptive community. The granularity of that snapshot (and the number of elided or revealed events) varies across metadata vocabularies. For example, a Dublin Core description, intended for relatively basic resource discovery, is a particularly coarse snapshot. A Dublin Core description of a postcard of the Mona Lisa might list Leonardo Da Vinci as the creator even though numerous events took place in between Da Vinci's creation and the representation of the Mona Lisa on a postcard. On the other hand, an INDECS description, for which the events associated with transfers of rights are extremely important, might describe more fine-grained event snapshots. Establishing the identity of the events implied in the respective snapshots makes it possible to associate descriptive properties in each metadata description with these events, which then facilitates mapping among properties in the metadata descriptions.

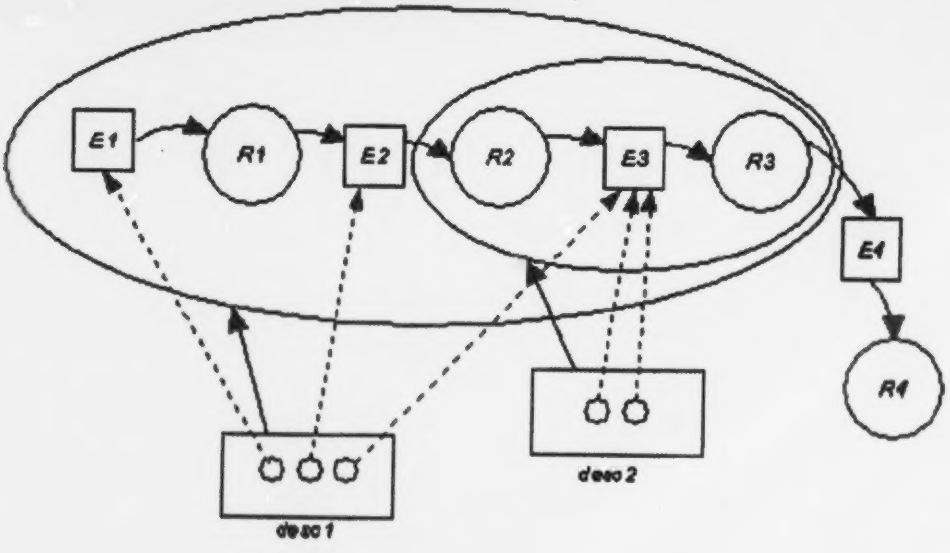


Figure 1 - Metadata and events

This basic concept of using events in mapping amongst metadata schema is illustrated in Figure 1. The larger circles represent manifestations of a resource as it moves through a set of event transitions; the events are represented by the squares interspersed between the circles. For example, event *E1* may be a creation event that produces resource *R1*. This resource may then be acted on by a translation event - event *E2* - producing resource *R2* and so on. The rectangles at the bottom of the figure represent metadata descriptions (instances of particular metadata vocabularies), and the ellipses that enclose part of the resource/event lifecycle represent the snapshot of the lifecycle addressed by that particular metadata description. For example, the larger dark-shaded ellipse represents the snapshot described by *desc1*, and the smaller light-shaded ellipse the snapshot described by *desc2*. The smaller circles within each descriptive record are the actual elements, or attributes, of the description. The dotted lines (and the color of each circle) indicate the linkage of the metadata element to an event - as shown the elements in *desc1* are actually associated with three different events that are implicit in the snapshot. For example, the attributes (moving from left to right) may describe *creator, translator,* and *publisher,* which are actually "agents" of the events. As shown, the three rose colored elements are all associated with a single event *E3*, implying a relationship between them that can be exploited in mapping between the two descriptive vocabularies that form the basis for the different descriptions.

The Nature of an Event Model

This paper has up to this point presented a number of justifications for the incorporation of event-awareness into the cataloging model. This section illustrates event-awareness by summarizing the modeling work in the Harmony Project. The full details of the Harmony work are out-of-scope for this paper. The interested reader should consult the research papers and reports [8, 9, 22] that provide greater details.

Over the last year, the Harmony Project has been examining a number of metadata vocabularies in an attempt to understand the entities and relationships that are common across them. The result is the so-called ABC model, which declares these entities as a set of base classes to which properties relevant to information content and its lifecycle can be attached. These entities are Resource (the primitive entity as it is defined in RDF), Event, Input, Output, Act (with Agent and Rols), and Context (with Date and Time). A UML representation [7] of the ABC model is shown in Figure 2.

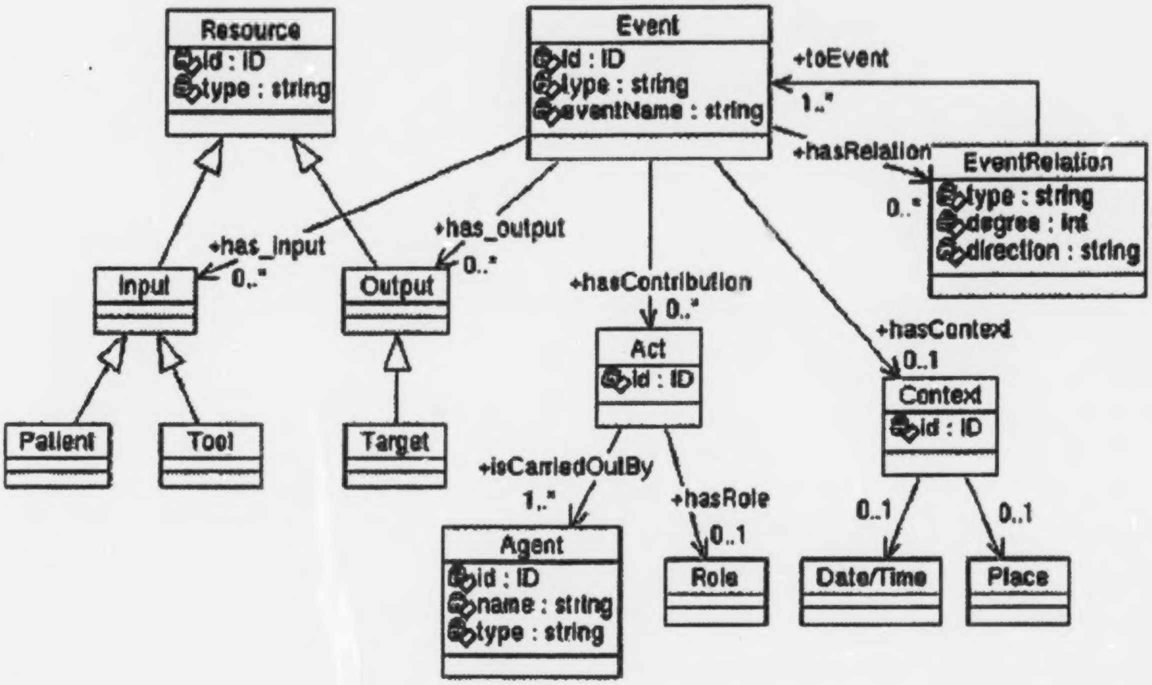

Resource
id : ID
type : string
Event
id : ID
type : string
eventName : string
+toEvent
1..*
+hasRelation
EventRelation
0..*
type : string
degree : int
direction : string
+has_input
0..*
+has_output
0..*
Input
Output
+hasContribution
0..*
+hasContext
0..1
Act
id : ID
Context
id : ID
Patient
Tool
Target
+isCarriedOutBy
1..*
+hasRole
0..1
0..1
0..1
Agent
id : ID
name : string
type : string
Role
Date/Time
Place

Figure 2 - UML representation of ABC model

We have tested and continue to refine this model in a number of experiments. For example, consider the following simple example of a digital audio:

The recorded performance was part of the "Live at Lincoln Center" series, made at The Lincoln Center for the Performing Arts on April 7, 1998 at 8PM Eastern time. The orchestra is the New York Philharmonic, and the musical score is "Concerto for Violin". The actual audio is a 130 minute MP3 encoding.

This example is represented in the ABC model in Figure 3 using UML-like symbols.

BEST COPY AVAILABLE

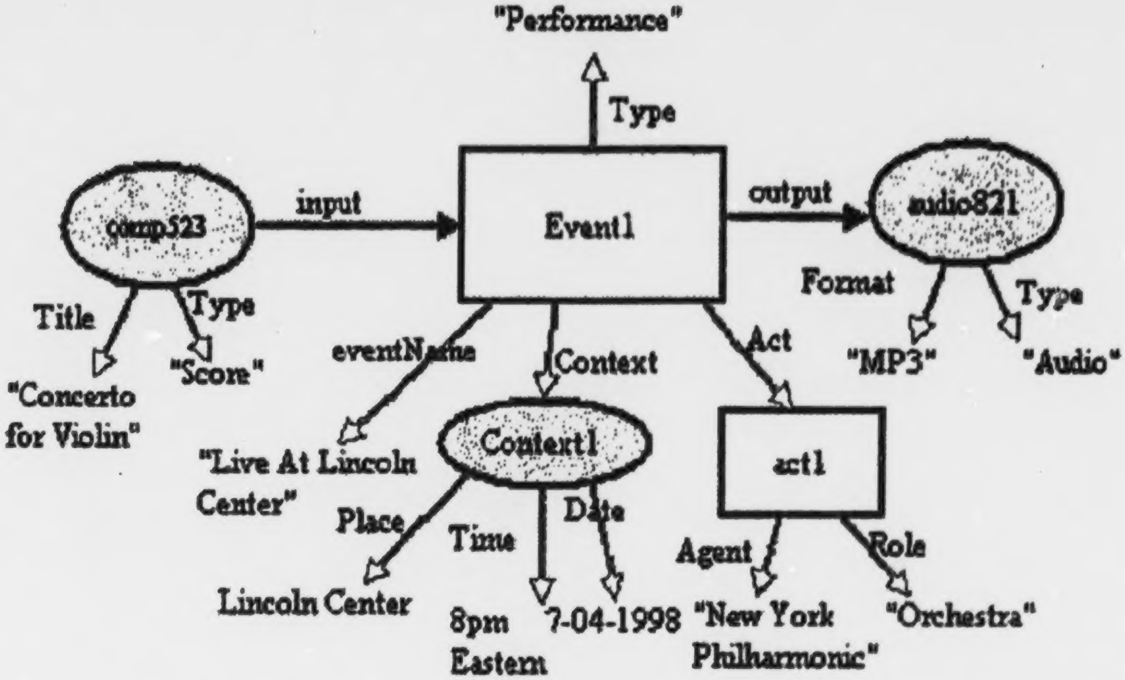

"Performance"
Type
comp523
input
Event1
output
audio821
Title
Type
"Concerto for Violin"
"Score"
eventName
Context
Act
Format
Type
"MP3"
"Audio"
"Live At Lincoln Center"
Context1
act1
Place
Time
Date
Agent
Role
Lincoln Center
8pm Eastern
7-04-1998
"New York Philharmonic"
"Orchestra"

Figure 3 - Example of ABC event-aware model

As illustrated in both Figure 2 and Figure 3, the model provides well-defined attachment points for various properties, by explicitly representing entities. Thus, the date of performance is defined as a property of the "performance" event, rather than as a property of the audio. The usefulness becomes clearer if we expand the example and include a "composition" event that feeds into "comp523" in Figure 3, with a "Date" property of 3-01-1804. This stands in contrast to a resource-centric model in which both dates (and perhaps) several others would be listed as cataloging properties of the single audio resource.

At this point we have experimented with the ABC model for mapping between a number of metadata schemas including Dublin Core, ID3 tags embedded in MP3, MPEG-7 descriptions in DDL, and the CIDOC CRM model. We have demonstrated that it is possible to do simple mappings using XML schema [6, 38] and XSLT [14]. The limitations of these tools has constrained the expressiveness of these mappings and in Harmony we are beginning to experiment with more powerful tools such as a metadata term ontology and the use of a general mapping rule language.

Conclusion

This paper has proposed radical changes in use of the catalog and the model upon which it rests. It has described why these changes are necessary if the library is to transition effectively into the digital age.

BEST COPY AVAILABLE

Changes of such magnitude obviously require careful consideration and strategic planning on the part of libraries and associated information professionals. They will require libraries to take a prominent role in research initiatives and, correspondingly, allocate resources to develop and hire the professionals capable of participating and leading such research. Being too conservative will only widen the disconnect between the rapidly changing information environment and the manner in which libraries profess to manage it. I end with an appropriate admonition from the NRS report [15] (taking the liberty to replace explicit references to "the Library of Congress" with "libraries"):

The alternative to progress along these lines is simple: [libraries] could become a book museum....But a library is not a book museum. A library's value lies in its vitality, in the way its collections grow, and in the way that growth is rewarded by the diverse and innovative uses to which its collections are put. [Libraries] will, by the choices [they] make now and in the next months and years, determine how much of that vitality will survive into the new millennium and how well [they] can avoid subsiding into diminished relevance.

Acknowledgements

This paper benefits from discussions and joint research with a number of people. The approximately 14 months spent on the NRC Library of Congress study were among the most valuable in my life and I thank all my colleagues there for their inspiring thinking. I owe special thanks to my colleagues in the Harmony project, especially Jane Hunter who has done wonderful work on metadata mapping using the Harmony ABC model. Thanks also to Clifford Lynch for referring me to the valuable papers form the CLIR authentication workshop. Finally, I express gratitude to the organizers of the workshop for inviting me and giving me the chance to think about these issues. Naomi Dushay also supplied invaluable editing advice. Support for work on this paper came from NSF Grant 9905955.

References

[1] *Dublin Core/MARC/GILS Crosswalk*, http://lcweb.loc.gov/marc/dccross.html.

[2] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions http://www.ifla.org/VII/s13/frbr/frbr.pdf, March 1998.

[3] W. Y. Arms, "Automated Digital Library: How Effectively Can Computers Be Used for the Skilld Tasks of Professional Librarianship?," *D-Lib Magazine*, 6 (7/9), http://www.dlib.org/dlib/july00/arms/07arms.html,, 2000.

[4] D. Bearman, "An Indefensible Bastion: Archives Repositories in the Electronic Age," Archives and Museum Informatics, Pittsburgh, Technical Report 13, 1991.

[5] D. Bearman and K. Sochats, "Metadata Requirements for Evidence.," Archives & Museum Informatics, University of Pittsburgh, School of Information Science, Pittsburgh, PA http://www.lis.pitt.edu/~nhprc/BACartic.html, 1996.

[6] P. V. Biron and A. Malhotra, "XML Schema Part 2: Datatypes," World Wide Consortium, W3C Working Draft WD-xmlschema-2-2000025, http://www.w3.org/TR/xmlschema-2/, April 7 2000.

[7] G. Booch, J. Rumbaugh, and I. Jacobson, *The unified modeling language user guide*. Reading Mass.: Addison-Wesley, 1999.

[8] D. Brickley and R. V. Guha, "Resource Description Framework (RDF) Schema Specification," World Wide Web Consortium, W3C Candidate Recommendation CR-rdf-schema-20000327, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/, March 27 2000.

[9] D. Brickley, J. Hunter, and C. Lagoze, "ABC: A Logical Model for Metadata Interoperability," Harmony Project, Working Paper http://www.ilrt.bris.ac.uk/discovery/harmony/docs/abc/abc_draft.html, 1999.

[10] M. K. Buckland, "What is a "document"?," *Journal of the American Society of Information Science*, 48 (9), 1997.

[11] P. P. S. Chen, *Entity-relationship approach : the use of ER concept in knowledge representation*. Washington, D.C.: IEEE CS Press, 1985.

[12] P. P. S. Chen, *The entity-relationship approach to logical database design*. Wellesley, Mass.: QED Information Sciences, 1991.

[13] C. M. Christensen, *The innovator's dilemma : when new technologies cause great firms to fail*. Boston, Mass.: Harvard Business School Press, 1997.

[14] J. Clark, "XSL Transformations (XSLT)," World Wide Web Consortium, W3C Recommendation REC-xslt-19991116, http://www.w3.org/TR/xslt, November 16 1999.

[15] Committee on Information Strategy for the Library of Congress, *LC21: A Digital Strategy for the Library of Congress (2000)*: National Academy Press, Washington, DC, 2000.

[16] Committee on Intellectual Property Rights in the Emerging Information Infrastructure, *The Digital Dilemma: Intellectual Property in the Information Age*. Washington, D.C.: National Academy Press, 2000.

[17] P. Conway, "The Relevance of Preservation in a Digital World," Northeast Document Conservation Center, Andover, MA http://www.nedcc.org/plam3/tleaf55.htm, February 1999.

[18] A. Crespo and H. Garcia-Molina, "Archival Storage for Digital Libraries," presented at Third ACM International Conference on Digital Libraries, Pittsburgh, PA, 1998.

[19] L. Duranti, *Diplomatics: New Uses for an Old Science*. Lanham, MD: Scarecrow Press, 1998.

[20] S. Harnad, "Free at Last: The Future of Peer-Reviewed Journals," *D-Lib Magazine*, 5 (12), http://www.dlib.org/dlib/december99/12harnad.html,, 1999.

[21] P. B. Hirtle, "Archival Authenticity in a Digital Age," presented at Authenticity in a Digital Environment, Washington, D.C., 2000.

[22] J. Hunter and D. James, "Application of an Event-Aware Metadata Model to an Online Oral History Archive," presented at ECDL 2000, Lisbon, 2000.

[23] ICOM/CIDOC Documentation Standards Group, *CIDOC Conceptual Reference Model*, http://www.ville-ge.ch/musinfo/cidoc/oomodel/.

[24] C. Lagoze, "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata," *D-Lib Magazine*, 2 (7/8),http://www.dlib.org/dlib/july96/lagoze/07lagoze.html,, July/August, 1996.

[25] O. Lassila and R. R. Swick, "Resource Description Framework: (RDF) Model and Syntax Specification," World Wide Web Consortium, W3C Proposed Recommendation PR-rdf-syntax-19990105, http://www.w3.org/TR/PR-rdf-syntax/, January 1999.

[26] D. Levy, "Cataloging in the Digital Order," presented at The Second Annual Conference on the Theory and Practice of Digital Libraries, 1995.

[27] D. M. Levy, "Fixed or Fluid? Document Stability and New Media," presented at 1994 European Conference on Hypermedia Technology, 1994.

[28] D. M. Levy, "Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment," presented at Authenticity in a Digital Environment, Washington, D.C., 2000.

[29] C. Lynch, "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis on the Central Role of Trust," presented at Authenticity in a Digital Environment, Washington, D.C., 2000.

[30] C. Lynch, "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information," *D-Lib Magazine*, 5 (9), http://www.dlib.org/dlib/september99/09lynch.html,, 1999, September.

[31] S. McKemmish, G. Acland, N. Ward, and B. Reed, "Describing Records in Context in the Continuum: the Australian Recordkeeping Metadata Schema," Monash University, Records Continuum Research Group http://www.sims.monash.edu.au/rcrg/publications/archiv01.htm, 1998.

[32] Metadata Ad Hoc Working Group, "Content Standard for Digital Geospatial Metadata," Federal Geographic Data Committee, Washington DC FGDC-STD-001-1998, http://www.fgdc.gov/standards/documents/standards/metadata/v2_0698.pdf, 1998.

[33] N. B. Olson, *Cataloging Internet Resources*. Dublin, OH: OCLC Online Computer Library Center, Inc., 1997.

[34] S. Payette and C. Lagoze, "Value-Added Surrogates for Distributed Content: Establishing a Virtual Control Zone," *D-Lib Magazine*, June , http://www.dlib.org/dlib/june00/payette/06payette.html,, 2000.

[35] G. Rust and M. Bide, "The INDECS Metadata Model," http://www.indecs.org/pdf/model3.pdf, July 1999 1999.

[36] F. B. Schneider and National Research Council (U.S.). Committee on Information Systems Trustworthiness, *Trust in cyberspace*. Washington, D.C.: National Academy Press, 1999.

[37] E. Svenonius, *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press, 2000.

[38] H. S. Thompson, D. Beech, M. Maloney, and N. Mendelsohn, "XML Schema Part 1: Structures," World Wide Web Consortium, W3C Working Draft WD-xmlschema-1-

2000225, http://www.w3.org/TR/xmlschema-1/, April 7 2000.

---

[1] Thanks to Stuart Weibel of OCLC for introducing me to the notion of metadata as a disruptive technology.

[2] Traditional cataloging is one form of metadata – a form of description or "data about data".

[3] This paper is not intended as a summarization of that report. Although this paper benefits from conversations during the NRC study, the thoughts and opinions expressed here are of the author alone.

[4] The interested reader may wish look at the many good sources of information on data modeling including the classic materials on E-R (entity relationship modeling) [11, 12], and the excellent work in various descriptive communities [2, 23].

[5] The museum metadata community [23] and archival metadata community [31] have recognized the importance of event-oriented models.

[6] Actually a URL is one of but several types of "locators" in the 856 field. For example, the contents may be a URN; a so-called permanent and location-independent identifier. While the permanence of a URN is an attractive concept, from the implementation point of view a URN is simply one or more levels of indirection to a URL, where permanence rests on the stability of the agency maintaining the indirection mechanisms. Moral: URNs really provide no real technical solution to the problems of fixity discussed here.

[7] For example, hashing techniques are generally insensitive to the difference between a trivial font change and a change in the wording of a paragraph.

[8] The community of people advocating this approach with David Bearman is collectively known as the "Pittsburgh Project".

[9] See explanations of the science of diplomatics in [19].

[10] Consider the highly descriptive FGDC standard for geospatial resources [32].

[11] Mapping among descriptive formats is not entirely new to the cataloging community. There have been numerous experiments with *crosswalks* between MARC-based cataloging records and metadata in its various forms [1]. These crosswalks generally presume a role where the catalog is the superior form

and other metadata forms have reduced functionality and, therefore, importance. This is different from the catalog acting as a mapping mechanism among distributed metadata packages that in their composite equally contribute to the "cataloging record" of a digital object.

[12] The result is a Warwick Framework-like container whose packages are distributed across multiple servers.

Library of Congress
Comments: lcweb@loc.gov (October 19, 2000)

- Conference Home Page
- What's new
- Greetings from the Director for Cataloging
- Topical discussion groups
- NAS study and 2 articles from the LC staff *Gazette*
- Conference program
- Speakers, commentators, and papers
- Conference sponsors
- Conference discussion list
- Logistical information for conference participants
- Conference Organizing Team
- Cataloging Directorate Home Page

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Linda Arret

Network Development and MARC
Standards Office
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4160

Carolyn Larson

Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4160

Descriptive Resource Needs from the Reference Perspective

BEST COPY AVAILABLE

Library of Congress Home Page

About the presenters:

Linda Arret is a network specialist in the Library of Congress Network Development and MARC Standards Office, where she focuses on issues related to reference and public services. Linda's experience as a frontline reference librarian has been instrumental in projects she has helped lead and plan, including online catalog development, public access to the Internet, public and staff training programs, reference presence on the Web, and collaborative efforts for providing digital reference services.

Carolyn Larson has worked for many years in reference at the Library of Congress, where she has been active on various automation committees relating to staff and public training, user interface design, and indexing issues. She is currently a business reference librarian in the Science, Technology and Business Division. In addition, she is a member of the Library's Bibliographic Enrichment Advisory Team (BEAT), a research and development team charged with the "development and implementation of initiatives to improve the tools, content, and access to bibliographic information," serving as Project Manager of the BECites+ Project as well as participating in the BEOnline+ Project. She is also a member of the ALA RUSA MARS Task Force on the Best of Free Reference Web Sites.

Full text of paper is available

Summary:

Drawing on the results of a survey to be conducted this summer, we plan to address the following topics from the perspective of reference providers:

Optimum "levels" of library and metadata descriptions (including descriptive/subject/administrative/access metadata) for content retrieval of Web-based resources (e.g. full MARC records; simpler, more structured Dublin Core records);

Descriptive needs that professional reference providers feel to be essential in performing their work (e.g. more subject data, more summary information);

Additional descriptive elements which reference librarians feel would facilitate achieving accurate and useful content retrieval in response to user queries and information demands;

Traditional concepts, such as authority files, uniform titles, specialized thesauri, that might be incorporated into metadata descriptions to facilitate resource discovery;

Problems, which might be addressed through improved interaction between metadata and present-day technologies, that arise as reference providers navigate the current "continuum" of resource discovery from catalog through "middleware tools" (such as pathfinders, finding aids, abstracting and indexing services, and databases) to content.

Descriptive Resource Needs from the Reference Perspective:
Report on a Survey of US Reference Librarians for the Bicentennial Conference on Bibliographic Control for the New Millennium

Library of Congress
November 15-17, 2000 by

Carolyn Larson
Business Reference Librarian
Science, Technology, & Business Division
Library of Congress
and
Linda Arret
Network Development Specialist
Marc Standards and Network Development Office
Library of Congress

Final version

This paper presents a discussion of what reference librarians require with regard to the bibliographic control of networked resources, based on 200 responses to a survey of U.S. reference providers and on comments made at an open meeting on this topic sponsored by the Library of Congress at the July 2000 American Library Association (ALA) Annual Meeting in Chicago.

Overall, the responses to the survey reflect the growing reliance of reference providers on Web-based resources. Almost half reported consulting print/microfilm resources and local networked digital resources "somewhat" or "much less frequently," than a year ago whereas considerably more than half reported consulting subscription Internet resources, search engines, and other freely available web-based resources "somewhat" or "much more frequently" than a year ago.

Approximately forty percent of the responding libraries reported providing access to **online subscription** or **selected free Internet** resources through the OPAC. Most of the remainder provide access only through web lists or book marks. In contrast, over eighty percent of the respondents indicated that, in their opinion, selected Internet-based resources **should be** included in the OPAC. Overall, respondents suggested that if Web-based resources are included in the OPAC, it would be most useful from a public service perspective if there were records in the catalog for the individual titles included within aggregated resources or within a Web site rather than for the aggregated resource itself or for the top-level Web site only. For those libraries cataloging either subscription or free Internet resources, about two-thirds report providing AACR2 level cataloging. With regard to the cataloging elements most needed for searching, over two-thirds of the respondents selected the following ten: title, subject keywords, URL, author/creator, link to an index/keyword search of the resource itself, controlled subject vocabulary, date of last update, time period covered by the resource, language of resource, and links to table of contents. With regard to those cataloging elements that must be included in the catalog record regardless of whether or not they are used for searching, over two thirds selected the following ten elements in addition to the previously listed ten elements: date of creation, genre, publisher, copyright/access restrictions, relationships to other works, format, geographic coverage, summaries of the resource, other unique identifying numbers and a statement that the resource is peer-reviewed.

In their free text comments, respondents singled out a number of problems with regard to the bibliographic control of Internet resources within the OPAC including: the need for title access for the full-text titles included in aggregator databases; the need to collapse multiple records for multiple versions/formats of the same intellectual content into a single "record" for public view; the need for greater collaboration both between public service and technical service departments within institutions and among multiple institutions in the selection and control of these resources; the need to ensure the long-term availability of those networked resources added to the OPAC's through greater attention to archival issues, and finally, the need to develop a single user-friendly "interface" that would allow users to link across relevant databases .

Respondents also included comments related to improving search retrieval on the Web at large. These included suggestions focusing on the use of intelligent agents, automated categorization of Web resources, information visualization technologies, and the **application** of concepts from traditional librarianship coupled with the use of XML and innovative technology, most notably in proposals to find ways to match natural language queries with standardized subjects and authorized names and in proposals for encouraging widespread use of unique identifiers for web pages or content including proposals to work cooperatively with selected publishers in order to provide **librarian-created** metadata to publishers which they could add to the HTML headers of their resources.

Survey Focus:

In developing the survey which forms the basis for our presentation, we attempted to address the following:

- Standard descriptive elements that professional reference providers believe to be essential in performing their work;
- Additional descriptive elements which reference librarians believe would facilitate achieving accurate and useful content retrieval in response to user queries and information demands;
- Optimum "levels" of library and metadata descriptions (including descriptive/subject/administrative/access metadata) for content retrieval of Web-based resources (e.g. full MARC records; simpler, more structured Dublin Core records);
- Traditional concepts, such as authority files, uniform titles, specialized thesauri, that might be incorporated into metadata descriptions to facilitate resource discovery;
- Problems which reference providers are experiencing in identifying relevant materials in an online environment which might be addressed through improved interaction between metadata and present-day technologies.

Description of survey instrument:

In an attempt to assess reference needs in this area, the survey contained:

- some basic questions about how the respondent's library was currently providing access to networked resources (and the reference librarian's satisfaction with current access at their institution). These questions focused on:
    - distinctions between bibliographic control of subscription and/or licensed resources and freely available Internet resources of "reference value"
    - the level of granularity of such bibliographic control and how libraries were coping with bibliographic control of resources supplied by aggregators and database producers.
- a series of questions focused on the cataloging elements reference librarians view as most important for inclusion in such records
- a series of questions focused on reference providers' reactions to various proposals that have been made for dealing with the cataloging of electronic resources, including:
    - providing access through separate web lists vs access through the OPAC.
    - providing different types or levels of control to subscription/licensed resources and "free" Internet resources
    - utilizing some type of "core" or "minimal" level of cataloging for electronic resources as opposed to full MARC cataloging.
    - providing single records vs multiple records for resources available in multiple formats (e.g. print and electronically).
- a number of open-ended questions about the major problems related to bibliographic control that reference librarians are facing in identifying (or assisting patrons in identifying) pertinent electronic resources and potential solutions to these problems from a reference point of view.

Finally, several questions were also included regarding the size and type of the library, the subject areas in which reference service is provided, the extent of use of electronic resources in providing that reference service, the number of years the respondent had worked in reference, and the respondent's general familiarity with various metadata schemes and projects making use of metadata. These later questions allowed us to consider if there might be differences among the answers given by groups of respondents based on the any of the above factors.

Background on the Respondents:

The Selection Process

Information about the survey (which was posted on the Library of Congress Web site) was sent by email to heads of reference or library directors of approximately 450 U.S. libraries. Drawing on unpublished statistics from the U.S. Dept. of Education showing number of reference transactions, size of staff, and expenditures by reporting U.S. libraries, information in the *American Library Directory*, and information on individual library web sites, we endeavored to contact a small, medium, and large public library and a small, medium, and large academic library in each US state[1], as well as a non-academic library from each of the Special Library Association chapters and divisions. In addition, we endeavored to contact each of the US state libraries and the four US national libraries. Each library was offered the opportunity to supply two responses. The survey was also announced at the end of June on a number of reference listservs prior to an open meeting at the July 2000 American Library Association Annual Meeting at which the topic was discussed, and twenty responses were obtained from that posting. A total of two hundred responses (representing one hundred sixty-nine libraries) were received, broken down as follows:

- Academic libraries: 70 responses (representing 58 libraries out of 159 libraries contacted)
- Public libraries: 56 responses (representing 47 institutions out of 151 libraries contacted)
- State libraries: 29 responses (representing 24 libraries out of 46 libraries contacted. Several state libraries were not contacted because a valid email address could not be located)
- Special libraries: 22 responses (representing 21 institutions out of 94 libraries contacted)
- US National libraries: 2 responses (Both from LC. The National Library of Medicine, the National Agricultural Library, and the National Education Library were also contacted, but did not respond. In addition, two of the responses in the Listserv category are from LC employees)
- Listserv postings: 20 responses (15 of those libraries identified themselves as "Academic," 2 as "public," 1 as "private non-profit"; 1 as "private for profit"; and 1 as "governmental.")
- Total responses: 200
- Total libraries represented: 169
- Total libraries contacted: 453

In addition, we received 17 direct replies (email or telephone) from librarians indicating that after looking at the survey, they felt that they could not respond. Their reasons varied from the academic librarian who said that between relocating and opening for the new school year, her library did not have the resources to respond, to the state library which indicated its function was more coordination than reference, to a special librarian, who wryly observed that the information center at her organization had been deep-sixed and she was now functioning in a different capacity within the organization. However, most of the libraries in this category gave as their reasons for non-response small size and lack of online databases of any description in their library. The following two comments are typical:

When I looked at the survey, I decided our library should pass the opportunity to respond on to another library from our area since in our small library we have neither an online catalog nor any databases.

When I got to the question about whether access to databases such as those available through FirstSearch was by OPAC or Web, I wanted to cry. I'd be happy if we had *ArtIndex* in electronic form at all, regardless of whether we accessed it from the OPAC or the Web, or both.

Characteristics of Respondents

A little more than half of the respondents (110) report they work in institutions having 1-10 reference staff. Slightly more than half (104) report working in reference less than fifteen years compared with those reporting fifteen or more years of work in reference positions (94); almost one third (61) reported more than twenty years experience. Thirty-eight respondents are part of what we are calling, for want of a better term, a "metadata aware" group, which includes those who indicated either "substantial knowledge" or "general understanding" of at least six out of eleven metadata projects/schemes listed on the survey or who indicated "substantial knowledge" of either three and "general understanding" of two of these projects/schemes or "substantial knowledge"of four and "general understanding" of one of them.[2] For the responding group as a whole, there was at least name recognition by approximately half of the respondents for five of the listed projects/schemes.[3]

A few of the respondents reported being in positions in which a single "reference" role dominated eighty percent or more of their time: "expert" end user, doing searches for patrons [24]; "trainer," teaching others to search (19); or "author," searching in order to prepare guides, current awareness services, training materials, etc. (1). As a group, however, most of the respondents indicated that their time was more evenly split between these three roles.

With regard to the subject areas most frequently searched, the responses indicate respondents are working in a wide variety of subject areas. Multiple responses were permitted on this question: a majority of respondents (118) selected "general reference", but there was also significant representation from the sciences, including medicine and technology (86), business (74), humanities (63), and arts (32); other areas mentioned by respondents included government documents, fire safety, newspapers and periodicals, current events and news, and local history and genealogy.

Their searching is primarily text-based, although somewhat less than a tenth did indicate that they spend up to one half their time searching for images. When using Web-based resources, the respondents reported being least likely to turn to such resources (subscription or free) to locate a specific fact, most likely to utilize them when searching for broad information on a particular topic.

Responses to several questions regarding the frequency with which these librarians are searching particular formats in connection with reference work, indicate that, overall, they spent the least amount of their time (under 20 percent) searching local networked digital resources; the remainder of their searching time as a group was fairly evenly distributed among print resources, subscription Internet resources, search engines, and online library catalogs.

Of perhaps particular interest to this conference are the responses to questions comparing how frequently the survey respondents are using various types of resources compared to a year ago. These responses reflect the growing reliance of reference providers on Web-based resources. Almost half report consulting print/microfilm resources (98) and local networked digital resources (90) "somewhat" or "much less frequently," whereas considerably more than half report consulting the following types of resources "somewhat" or "much more frequently": subscription Internet resources (142), search engines (122) and other freely available web-based resources exclusive of search engines and OPAC's (115).[4]

The Current Situation:

How libraries are providing control for Web resources

Karen Calhoun, in her paper for this conference [5], found all seven major ARL libraries which she surveyed were providing discovery and access to selected subscription resources both through the OPAC and via Library-created Web lists. Our survey results suggest a somewhat greater split among our respondents: sixty-eight libraries report access to subscription electronic resources through the OPAC;[6] ninety-seven through web lists alone; (with four libraries leaving the question blank).

Of those sixty-eight libraries reporting access to subscription Web resources through the OPAC, fifty-seven are cataloging the resource (such as *FirstSearch*) itself; forty report cataloging individual databases within resources such as *FirstSearch*, (for example, "Readers' Guide Abstracts"), whereas only thirty report cataloging full-text titles within such databases.

We also asked our respondents about the level of cataloging provided for these subscription Web resources. Bear in mind that the responses to these questions are those of reference librarians, not catalogers, who may or may not have consulted with catalogers at their institutions before responding. With this caveat in mind, around two-thirds report full level AACR2 cataloging for these resources, whether at the resource, database level, or individual title level, and the remainder indicated either "some other level" of cataloging or "not sure." This latter group was asked to select from a list those cataloging elements which they typically found in catalog records for subscription Web resources at their institution today. Elements (in order of frequency cited) selected by a majority of those who reported that their institutions catalogued either the resource or the database included: title, publisher, place of publication, URL, author/creator, format [7], date of creation of resource, genre[8], and time period covered by the resource. For those reporting their institution cataloged the individual titles within resources, only title, URL, publisher, format and time period covered were cited by more than half the respondents.

The situation with regard to free Internet resources showed a somewhat greater split. Only fifty-one of the surveyed libraries report adding free Internet resources to their OPAC's. All of them are also creating weblists of selected free Internet resources. An additional one hundred libraries are providing guided access to selected web pages only through bookmarks or web lists; thirteen libraries indicated they were neither developing webliographies nor adding records for such resources to their OPACS, and five left the question blank.

Table 1
Library Access to Web Resources

	Access via Web Lists Only	Access via OPAC Only	Access via Both Web Lists and OPAC
Subscription Web Resources	97	2	66
Selected Free Web Resources	100	0	51

With regard to the levels of cataloging of free Internet resources, of the fifty-one libraries reporting access through the OPAC, thirty-three indicated that these free sites are given full AACR2 cataloging, two reported such sites were cataloged at a Dublin Core level; eleven reported "some other level" or "not sure," and five left the question blank. We asked the group reporting "some other level" or "not sure," in a follow-up question, to select from a list, those cataloging elements they typically found in catalog records for free Internet resources at their institutions. Elements (in order of frequency cited) selected by a majority of those responding were: title, URL, author/creator, publisher, and place of publication. In other words, a shorter, but similar, list compared to the elements selected by this group for subscription web resources.

Degree of Satisfaction with the Current Situation

Somewhat surprisingly, by an overwhelming majority (144 of the 195 who answered this question), seventy-four percent indicated that the current method of access for web-based subscription resources at their institution (whether by Web, OPAC or both) was satisfactory for their work as reference providers.

Looking more closely at some of the characteristics of this group -- such as time spent accessing subscription web resources, type of library setting, number of reference staff, years of reference experience, subjects most frequently searched, and level of metadata awareness, we found that the percentage reporting satisfactory access remained fairly constant in all cases with the exception of those respondents working in institutions with more than twenty reference staff. In those cases, the percentage reporting satisfaction with the current mode of access dropped to just over fifty percent.

We might note that somewhat less than half of this satisfied group (60) are providing access to subscription Internet resources through the OPAC, and all that do provide catalog access also report access through web lists as well. Looking at these sixty who are accessing these resources from the OPAC, forty-seven report that the institution is cataloging the online resource itself, [9] thirty-three, or just over half, report that their library is cataloging at the database level,[10] while twenty-nine report that the OPAC contains records for the individual titles within such databases.[11] Twenty report that their institution is cataloging at all three levels.

Turning to free Internet resources, a smaller number, (114 out of 191), but still a majority responded positively to a similar question about satisfaction with access to free Internet resources. Of this group, sixty-nine are accessing these resources only from the Web; forty from both the web and the OPAC, one from the OPAC only; while four reported providing no guided access from either the OPAC or the Web.

Given the relatively large number of respondents who reported that the current situation is "satisfactory" for their reference work, it might be tempting to assume that there is nothing more to say and conclude the paper right here. However, if we look at responses in two other sections of the survey -- a series of questions on whether such resources **should** be in the OPAC and a series of free text comments on problems and wished-for improvements -- we discover some interesting things.

First, let us look at the responses to the survey questions on whether web resources should be in the OPAC. In line with some of the discussion on the Alternative Architecture thread on the Conference listserv, eleven of our survey respondents (out of one hundred seventy-five responding) felt that neither subscription nor free Internet resources should be added to the OPAC. An additional twenty-six would incorporate records for subscription Internet resources in the OPAC, but exclude records for free Internet resources. A few added comments to reinforce their position.

> Incorporating links to Internet resources from the catalog may open up the catalog to unrestricted Internet browsing which conflicts with Library Board policy, and may also result in lack of access to the catalog if the limited number of work stations are tied up by Internet users.
>
> I have always had some misgivings about offering access to all types of resources at the same time.....We have found it very useful to suggest students keep the idea of the in-hand materials and the method of locating them (the PACs) separate from the virtual web-based resources (periodical databases, Internet databases, etc.) and the more complex methods of searching them.
>
> In some ways, I think the library catalog should be restricted to materials the library actively acquires. Otherwise it is in danger of losing its identity. But I think some kind of cross-reference to other resources would be good.

However, the majority of respondents, including those who also reported that they found their current access satisfactory, came down on the side of adding such resources to the OPAC. Of the eighty-four who reported both that present access is satisfactory and that their institution provides access to subscription Internet resources only through web lists, forty-nine answered "yes" to the question on adding web-based subscription resources to the catalog. When asked if it would facilitate access to **individual titles** within subscription databases if records for them were added to the OPAC, the number of positive responses among this satisfied group, rose to sixty-six.

With regard to free Internet resources, thirty-six of the sixty-seven "satisfied" respondents who are not currently providing access to selected free Internet resources via their OPAC, answered "yes" to the question of whether it would facilitate access to these resources if they were added to the online catalog."

Looking at the comments of those who reported that access is satisfactory for their work as reference providers, we find additional evidence that these providers still see the need for improvements in that access. Among those who report their current access is satisfactory and who provide access through the OPAC, we found strong comments from respondents in institutions which do not catalog individual full text titles supplied by aggregators indicating a need for such access, a point even more forcefully brought home by respondents who indicated that the current situation at their institution was **not** satisfactory for their work as reference librarians. Both groups also cited difficulty in determining which journal titles are indexed in which online resources; while subject access for e-journal titles was viewed by others as inadequate. Several respondents pointed to technical problems arising in the OPAC when a single catalog serves a consortium, but access rights or access methods (IP address versus passwords) to individual resources vary by member institution.

On the other hand, among those who found access "satisfactory" but who provide access only through the Web, we found comments such as the following:

> I think it is important to let customers know if there is an Internet-based version of something in the online catalog. We should let them know that we have the New York Times in hard copy for a certain period, and also on microfilm, and also full-text through a subscription database. Customers should know that they can read the Ohio Revised Code in our Reference section, but that they can also access the text from the Ohio State website.

Among the reasons for dissatisfaction cited by users who access these resources only via the Web:

- problems with users and librarians finding or remembering what resources are available:
  - Although we feature the resources in many ways, they tend to get buried and lose their importance on our webpage.
  - It is necessary to hold too much information in one's head, that is, to remember all the places (we've already paid for) which might yield pertinent information for the question at hand.
- familiar complaints about maintenance and redundancy,
  - Maintaining lists of links ...on the library web page leads to the need for creating redundant links on the multiple subject guide pages libraries have gotten in the practice of developing. For example, a good biography site probably belongs on every subject discipline page .... but whether multiple people maintain the multiple subject guides or a single person maintains the subject site, every time a new site is added, if it is appropriate for all the subject pages, it has to be added physically to each.
- And a reminder of the primacy of the OPAC as a starting point for research which should be encouraging to this audience:
  - Users are still using the catalog systematically. If these resources are not in the catalog, they are not enough used.
  - Our online catalog should serve as a comprehensive record of all our resources, regardless of format, so librarians and patrons can tell what we have by looking in one place.

Descriptive needs professional reference providers feel to be essential

With regard to the cataloging elements reference librarians view as most important for inclusion in catalog records, the survey attempted to address this issue through questions on current use of various metadata elements by reference providers for searching, and their perceived need for the same elements either displayed on the catalog record or as search elements. When the question was phrased as cataloging elements needed for searching only two emerged as being currently used at least fifty percent or more of the time by two-thirds or more of the 191 responding to this question: title (138) and subject keywords(138) . Rephrasing the question to ask what they thought they would use if it were possible to make all of the elements were available to them as searching elements resulted in a considerably expanded list: Of the 196 responding to this question, two-thirds or more indicated the following would be "essential," or "often useful."

Table 2

Cataloging Elements Considered Essential or Often Useful for Searching

Cataloging Elements	**Total Respondents (out of 196 responding)**
Title	185
Subject: Keywords	178
URL	166
Author/Creator	165
Index/Keyword search of resource	159
Subject: controlled vocabulary	148
Date of last update	143
Time period covered by resource	135
Language of resource	133
Table of contents	132

A specific question distinguishing between the need for elements to be present in the catalog record even if they were not generally used for searching was added following discussion at the open meeting at the American Library Association Annual Meeting **where there was general agreement that all of the elements listed in question 32 of the survey need to be present on the catalog record.**[12] As one survey respondent said:

> While I may not search by each of the elements listed, they are all occasionally essential in that they provide information by which to evaluate the usefulness of the information relative to my need.

The following elements were selected as being either "essential" or "often useful" to display in the catalog record by two-thirds or more of those responding to this question.

Table 3

Cataloging Elements Considered Essential or Often Useful for Display

Cataloging Elements	**Total respondents (out of 167 responses)**
Title	166
URL	163
Date of last update	160
Author/Creator	157
Subject keywords	157
Language	156
Time period covered by resource	156
Date resource created	150
Genre	144
Publisher	141
Copyright/access rights	141
Subject controlled vocabulary	141
Relation to other works	140
Link to index/keyword search of resource	138
Format	138
Geographic coverage	134
Summary of resource (by librarian)	130
Other identifying numbers (e.g. ISSN, GPO)	129
Link to Tables of contents	126
Summary of resource (by publisher)	119
Statement that resource is peer-reviewed [13]	112

Indeed, we might note that each element on the survey was considered "essential" by at least some respondents; the element that was selected as "essential" the fewest number of times was a " link to or excerpt of a review of the resource" (10 responses). The element selected most often as "not important" was "subject classification code" (23 responses).

Elements which were selected as "essential" or "often useful" by less than two-thirds of the respondents included:

- excerpts or links to reviews of the resource (95)
- place of publication (90)
- subject classification codes such as LC or Dewey (79)

One thing that was somewhat surprising to us in looking at the above lists was the relatively low ranking of controlled subject vocabulary[14] compared to subject keywords, particularly since the need for controlled vocabulary was emphasized in the free text comments of just under a quarter of the respondents. In general, we found few differences among the responses in this regard given by librarians by type or size of library, those having more years service as reference providers, with the percentages of respondents rating controlled vocabulary as either essential or often useful hovering around two-thirds for most of these categories. The most pronounced differences appeared among those eighty-six respondents doing concentrated searching in the sciences and technology, where only fifty-three percent indicated controlled subject vocabulary was "essential" or "often useful"; the forty federal and state governmental librarians where eighty-two percent ranked "controlled subject vocabulary as "essential" or "often useful," and the "metadata aware" group of respondents, in which ninety percent of respondents in this category selected controlled subjects as "essential" or "often useful."

The comments of a few who did not rate controlled subject vocabulary as essential suggest that for some at least, it may be a matter of settling for what they view as practical.

Just getting a brief title and keyword record into the catalog would be better than it is now.

I would like speedy cataloging with minimal information.

Several other comments suggested that controlled subject headings broadly applied to resources as a whole cannot make the fine distinctions needed by patrons to focus in on their specific subjects:

And even if we catalog the web site, will the subject headings assigned be extensive enough to make it clear that the "Voice of the Shuttle: [Web page for Humanities Research" (http://vos.ucsb.edu/)] is a good place to go for links to William Blake, and any other individual author for that matter? This is doubtful. A major search engine would pick this up though.

Additional descriptive elements cited by reference providers

Respondents were also given the opportunity to list additional cataloging elements which they considered to be essential, for searching, or for viewing on the record, or for accessing directly from the record.[15] In some cases respondents expanded on the kinds of information needed in relation to elements in included in our list and we include that information here as well.

Table 4
Additional Cataloging Elements Cited by Respondents as Essential

Type	**Cataloging Elements**
Controlled vocabulary (in addition to controlled subject headings)	**Authorized names**
	Uniform Titles
Title Information	**Previous titles** -Include information on previous titles

	Variant titles - Include popular title added entries "for many web resources the title is often difficult to determine due to graphic elements that may or may not be part of the title."
Responsibility for resource	**Publisher**
	- record changes. For example if a Dept of Commerce publication moves to the private sector
	- indicate "authority" of publication; distinguish official site/version from a copy
	Author/Creator
	- include all authors - especially important for multiple authors of scholarly publications and
	- include affiliation
	- include residency/citizenship (important for localities trying to maintain a record of the intellectual output of a particular region.)
	- include name changes, especially corporate names, with links between old and new versions
Access Information	**Notes/Terms of availability**
	- include information on mirror sites
	- include subscription status (free/registration required/fee/pay-per-view)
	- pertinent access information for multiple campuses when the rights to content or access methods vary by campus
	- address of content provider or distributor
	- other access restrictions
	- handicap accessibility
	- specify if material is classified and include contact information for classifying agency
	- specify if material is encrypted and include information on encryption standards used.
Additional Dates	-update schedule of resource, if known
	- date catalog record was updated
Standard Numbers	- include PURL's
	- always include ISSN's, if available. Being used by libraries to link to print holdings for earlier volumes
	- develop a standardized system of digital identifier's so each web page catalogued has a unique number "so when URL changes identifier remains the same."
	- industry codes like NAICS so there can be cross-linking between OPAC and business databases that include these numbers in their metadata.

Resource Description	- presence of images, sounds, text, data, graphics, video clips, etc.
	- size of document/site; downloading information
	- formats available (e.g. .pdf)
Relation to other works	- specify if content is full text or full image. Note if portions of "full text" resources actually omit certain types of materials. (A ds, graphics, etc)
	- note where indexed - including non-electronic indexes.
	- enhancement of online version (multimedia, reference linking, etc.)

Optimum "levels" of library and metadata descriptions:

With regard to the "level" of cataloging needed for these resources, it would seem that the above results indicate that a majority of those responding see a need for description that could be accommodated at the level provided by a "Dublin Core"-like model. All of the elements listed in tables 1 and 2 can be accommodated by Dublin Core[16] and its qualifiers to distinguish between elements that may be repeated with distinct types of data: such as date created and date updated; subject keywords and controlled subject vocabulary; various identifiers such as URL's, ISSN's and descriptions containing tables of contents versus those containing summaries or other descriptive information. Looking at the elements which were added by respondents as "essential," we see a number that could be accommodated by Dublin Core, as well.

These results also appear to be consistent with Lundgren and Simpson's survey of graduate students regarding their need for various cataloging elements for the description on Internet resources.[17] In that study title, primary author, Internet address, and summary were ranked highest, followed by subject, level of information, titles of related works in print, date created, date updated, access limits, additional authors, recommended software, system requirements, size of file, edition, frequency, and inclusion of graphics.

Print and Online: One Record or Two

Another section of the survey which produced many, many comments related to the problem highlighted by Michael Kaplan [18] in his paper for the conference of single versus multiple records for works appearing in multiple formats: the following librarian perhaps put it most forcefully:

Multiple formats for a single serial title result in much confusion for patrons. Many give up rather than search through multiple records to find what they need .... there is great need for consolidating access to different versions or formats for serial titles. Our catalogers are "purists" and want a clean database that will migrate well, but this does not make for a user friendly catalog. We often see 8 to 10 record[s] for the same title: microfilm, microfiche, microcard, online, paper, title changes, etc., etc. Please help!

Trying to balance the strong public service voice for a single record for multiple formats with the technological realities of computer-to computer data interchange of aggregator-supplied data so well described by Kaplan (a method which appears to offer some solution to the equally strong public service call for help in supplying title access in the OPAC to aggregator-supplied titles[19]), a number of our respondents urged that someone find a way, as one put it, to: figure out some way to maintain a clean database to make the catalogers happy and make it user friendly for patrons. Be able to hang multiple formats and holdings information on a single record."

There should be a way for local control of "holdings" within the authority [master] record

Thus, we strongly second the similar suggestion made by Kaplan and urge that an effort be made to develop a means by which records can be merged "for the public view that are kept separate in the technical services components of our catalogs."[20]

Improved interaction between metadata and present-day technologies

Survey-respondents were also asked to respond to an open-ended question regarding problems which might be addressed through improved interaction between metadata and present-day technologies. Before looking at some of the specific problems raised and the suggestions which our respondents made, we would like to return to the question raised by the Alternative Architecture thread of the conference listserv. Should these resources be in the OPAC at all? Even as reference librarians pointed out over and over again some of the problems of adding Internet resources to the OPAC cited in Barbara Baruth's *American Libraries* article,[21] the majority also made it that they thought selected Internet resources had their place in the OPAC. However, looking at many of the comments supporting the inclusion of these resources in the OPAC, we see one theme emerging over and over again: a unified search interface that is clear and easy to use.

> Again, having one place to search that would include relevant resources would make research less fragmented.
>
> In my opinion, it would be better to go to one place (the OPAC) for all resources, rather than to have to search the OPAC for other materials and then hunt for Internet resources by browsing through extensive "webliographies."
>
> I'd rather have our users have a seamless way of searching for information...
>
> [The interface for accessing networked resources is a problem because of] lack of ease of use. Simple, easy and familiar are VERY critical and are not there yet.
>
> [Access to free Internet resources is a problem because] they are scattered over a variety of access points. For example, some in bookmarks, some in classroom handouts, others in general subject pages.

We bring these examples up because some of the solutions offered by our respondents actually suggest a more distributed approach, which might well be consistent with Barbara Baruth's observation: "The future of library systems architecture rests in the development of umbrella software that digests search results from rapid, coordinated searches of a variety of disparate databases-OPACs of locally-held print and audiovisual materials, union catalogs, consortial catalogs of e-books and journals, and specialized digital library collections."[22]

This sounds similar to us to Kaplan's observation that "we are beginning to see the development of search interfaces and search engines that will simultaneously search and unify heterogeneous databases consisting of MARC records in all the bibliographic formats, as well as databases comprising all present and future data structures: Dublin Core, EAD, TEI, CIMI, VRA Core, MAB, MAB2, etc.,"[23] and Calhoun's discussion of "the need to be able to manage loosely federated data from many sources."[24]

We might note in passing that the concept of using "simultaneous" automated searches of multiple databases is under active consideration by the LC-led Collaborative Digital Reference Service (CDRS) project.[25] A recent proposal by Donna Dinberg (National Library of Canada) to the DigiRef team, would make use of captured metadata from bibliographic citations included in responses to users' inquiries to generate automatic searches of appropriate catalogs and databases through the use of Z39.50 protocols, ultimately allowing users to determine local availability of the materials cited and to initiate a loan or document delivery request, if desired.[26] Comments from some of our respondents include:

> [I] would like a more "relational" approach. Example, a subject search would bring up a result that includes categories such as 1. [Library Name] Materials, 2. Free Internet Resources, 3. Fee-Based Online Resources, etc.
>
> [There is a need for] unified searching; cross-platform searching.
>
> We need more automated linages between our Web guides and our online catalog.
>
> .... If the descriptions are part of a federated system where the builders of discovery and access mechanisms have no control over metadata content, there may still be some value in building a search tool that propagates multiple search terms from a single term based on the term variants or equivalents listed in one or more established structures such as authority files or thesauri ... In the best of all possible RDF worlds, a discovery tool should be able to convert a user input search term into the normative form of the term for a given set of metadata, based on the coding of the description itself and offer manipulation of the full panoply of syndetic relationships offered by that particular normative form upon request.
>
> Map specialized vocabularies at the highest level of the hierarchy, e.g. Transportation thesaurus would be mapped to LCSH for general transportation terms such as "Transportation," "Vehicles, etc." the user would be alerted to the fact that the data received is from a non-LC collection, in this case the Dept. of Transportation data. The user would be given the option to enter into the specialized database on transportation or remain in the LC online catalog to continue a search.

Many other comments, which may or may not be applicable to searching distributed databases, clearly move beyond manual cataloging of Internet resources in the OPAC to talk about incorporating various automated solutions into the mix. Such comments focused on some of the following problems:

On the "disappearing" web:

- Perhaps indexed metadata might include "control numbers" keyed to URL/path so when harvesters build indexes, any metadata value that returns a 404 error produces a mechanism for that search engine to remove all index terms built from metadata in that control number.
- Retrieve "close matches" when 404 error occurs, possibly by truncating the search to the "root" URL
- Additional use of URL and content checking software; perhaps with automated notification service - or even better - automated updating of the metadata record.
- More use of unique identifiers for identification of electronic resources; greater use of PURLs or similar technology plus greater use of automated link checking tools.
- Archive resources "of research value"; provide "traditional" cataloging only for archived resources.

On Copyright issues:

- Set up a system like music played on the radio where any station can play anything and have a formula whereby copyright holders get paid reasonably, but there is little administrative burden on the patron
- For issues of copyright, see Mike O'Donnell's paper presented at the May 2000 National Online meeting, where he discusses a proposal for an IP (Intellectual Property) meter. Clicking on (c) symbol would show the information needed to license the content, including "who owns the material and who publishes the material," along with the "permissions, i.e., "how the content may be used and what it will cost, ...[and] the terms of use."[27] All of this is managed by a copyright clearing house which collects any fees, sends the payments to the publisher with info on who licensed the content, and supplies content to purchaser with a digital marking showing the original source.

Enhanced searching

- Enhance existing search engine technology to move beyond matching strings of characters to search concepts or meaning as well through techniques such as disambiguation, contextual and grammatical parsing and use of semantic networks to increase precision.
- The ability to fine-tune a search, using such methods as frequency of words or location of words and element identification (such as author, title)
- ... better or more exacting search engines within huge searchable databases [would be useful]. I'm not speaking of big search engines like Google, but using this kind of power in smaller subsets like newspaper archives, etc.
- Study how nonlibrarian users do their searching.

On use of metatags:

- Require records management/life cycle controls (metadata fields) on anything added to a library catalog or web page. Unfortunately, this means you'd lose a lot of good information offered for free but if pushed for in quality publishing circles, [this] might be a selection criteria that might drive the market...
- Issuing agencies should include in their metadata the agency name, using a standardized format and they should use controlled subject headings to describe the resource.
- Use XML to create specialized tags for necessary data such as author and title. However, I would also want the metadata to be visible to the user so that the information could be used to construct better searches later.
- If metatag information was more rule-based, as in standard library cataloging, changes in resource URL's would be less of a problem. Searches on resources that use consistent metadata would have repeatable results regardless of some types of changes now inherent in the web.
- Work with selected publishers providing them with **librarian-created** metadata which they could add to the headers of their resources.

On controlled vocabulary:

- Ideally it would be beneficial if the "natural language" type of search could somehow be automatically screened through an "authority file" to eliminate large numbers of false drops. For example, I want material on housebreaking my dog. If as a patron I could type in housebreaking dogs and receive only results on "dogs- habits & behaviors" (or whatever the current "official subject heading" is) I would probably be happier with the result of my search query.

Among the most expansive and thorough responses to this question were those provided by Gerry McKiernan, Science and Technology Librarian at Iowa State University, curator of Cyberstacks, and field editor for the *Journal of Internet Cataloging*. His response to this question on the survey was a referral to his 1999 article in the *Journal of Internet Cataloging*[28] and to related resources cited on his web page.[29] In these sources McKiernan refers to a number of projects to facilitate access to relevant Web resources including the use of intelligent agents,[30] automated categorization of Web resources, such as OCLC's Scorpion technology used in its CORC project[31]; information visualization technologies such as the SPIRE(tm) suite of information access and visual analysis tools developed by the Pacific Northwest National Laboratory;[32] and natural language processing programs such as DR. LINK and KNOW-IT of MNIS-TextWise Labs.[33]

We believe that what the above responses suggest is that many public service staff recognize, as Michael Kaplan said in his paper for this conference, that we *are* "really, really drowning" in the sea of electronic resources[34], and carefully hand-crafted records for each, is an impossible dream. We need whatever help technology can give us, and we as a profession need to maintain awareness of the possibilities of current research and openly communicate and work with researchers in these areas if these problems are to be solved.

Concluding thought

In conclusion, we might note another theme that surfaced both at the ALA 2000 meeting and in the comments portion of the survey relating to the interaction of public service and technical services staff and departments. On the one hand there was clear evidence of the walls that have grown between reference and technical services, as reflected in the following comments:

> It's unfortunate that the structure of our libraries into public and technical services units inhibits communications between reference librarians and cataloguers. This survey should not be about "my wishes." It should be about a real conversation that has to go on in the library.

On the other hand, there were also signs of the breakdown of those walls in some places, which librarians described both at the open meeting at the American Library Association Annual Meeting in July 2000 and in comments on the survey. Whether through participation in CORC[35] or through other homegrown efforts at collaborative work between public and technical services in the selection and bibliographic control of web resources[36], these projects easily fit within the framework described by Karen Calhoun in her paper for this conference of the "typical" progress of a new electronic resource through the "resource description" process at many institutions.[37] Even where there was no collaboration mentioned, there were many comments calling for increased communication between the two departments. In this regard, we found our survey results to be very consistent with proposals put forward by Karen Calhoun on the redesign of library workflows within institutions, making increased use of cross-functional virtual teams for the selection and cataloging of networked resources.

Finally, we would like to point out one additional characteristic of our respondents which we think speaks directly to this point and clearly reflects the importance of the topic of this conference to reference providers. In response to the concluding question on the survey asking respondents to leave their email address or other contact information if they wished to receive additional information about the results of the survey or the Bicentennial Conference, one hundred and nine respondents did so. When was the last time you asked a survey-taker to keep in touch? Clearly, there is interest and concern about this topic in the reference community.

Notes:

1. Because of incomplete and inconsistent data for some institutions along with variations in the number and size of institutions in each state, "small," "medium," and "large" were loosely defined, and varied somewhat from state to state and by type of library. Generally, we began by looking at the U.S. D.O.E. data for each state and trying to select one institution with numbers falling in the top quarter, one from the bottom quarter, and one from the midrange of the data available for that particular state, supplementing this information with other sources as necessary.
2. Respondents were asked to rank their knowledge of the following according to these choices: "Have used or have substantial knowledge," "Have a general understanding," "Recognize the name only," or "Have never heard of it": Extensible Markup Language (XML), Standard Generalized

Markup Language (SGML), Resource Description Framework (RDF), Dublin Core (DC), Digital Object Identifiers (DOI), Text encoding Initiative, (TEI), Encoded Archival Description (EAD), Cooperative Online Resource Catalog (CORC), Consortium for the computer Interchange of Museum Information (CIMI), Scout Report Signpost, and Jointly Administered Knowledge Environment (jake).

3. CORC , XML, SGML, Dublin Core, and Scout Signpost.
4. Only for the category "online library catalogs" did a majority report that they consulted them "about the same" as a year ago.
5. Calhoun, Karen, "Redesign of Traditional Library Workflows: Experimental Models for Electronic Resource Description," Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000. (http://lcweb.loc.gov/catdir/bibcontrol/calhoun_paper.html) (6 Dec 2000) (hereafter cited as Redesign)
6. Sixty-five of these respondents also report access directly from Web pages is also available.
7. The survey question contained the following explanatory note: "Format: (including description of the software, hardware, or other equipment needed to display or operate the resource)."
8. The survey question contained the following explanatory note: "Resource types (genre). For example: abstracting/indexing services, working papers, technical reports, dictionaries."
9. These respondents responded positively to the question: "For example, if your library subscribes to OCLC's *FirstSearch,* is there a record in the online catalog for *FirstSearch*?"
10. These respondents responded positively to the question: "For example, if your library subscribes to OCLC's *FirstSearch*, are there records in the online catalog for individual Firstsearch databases that are part of your subscription, such as ReadGuid Abs(Readers' Guide Abstracts), HumanitiesAbs (Humanities Abstracts), PAISIntl (Public Affairs Information Service International), etc.?"
11. These respondents responded positively to the question: "For example, if your library subscribes to the database "Periodical Abstracts with full Text" in FirstSearch, are there records in the catalog for the online versions of the journal titles indexed in that database?"
12. Title, Author/Creator, Publisher, Place of publication, Date of creation, Date of last update, Resource type: genre, Format, Copyright restrictions, Relation to other works and formats, URL, Other unique identifying numbers or codes, Time period covered by the resource, Language of the resource, Subject: controlled vocabulary, Subject: keyword, Subject: classification code, Summary/annotation of the resource (publisher supplied); Summary/annotation of the resource (librarian supplied); table of contents; links to index or keyword search of the resource, excerpts or links to reviews of the resource; information that the resource has been "peer reviewed."
13. At the LC-sponsored open meeting on this topic at the American Library Association Annual Meeting (July 2000) several college and university librarians stressed that information indicating that articles in a title are "peer-reviewed" is indispensable in an academic setting; and indeed, among those respondents from academic institutions, just under two thirds (56) indicated this piece of information was either essential or often useful
14. Eighty-four percent of the respondents indicated that it is essential or "often useful" for controlled subject to appear on the catalog record for networked resources compared to ninety-four percent selecting "essential" or "often useful" for subject keywords.
15. As one respondent said: "I am assuming that ... clicking onto one of the elements [in the bibliographic description] will take you to a page that will have many of the elements I marked as being 'not essential.'

16. The Dublin Core Metadata Initiative. (http://purl.org/DC/index.htm) (6 Dec 2000)
17. Lundgren, Jimmie and Betsy Simpson. "Looking Through Users' Eyes: What Do Graduate Students Need to Know About Internet Resources via the Library Catalog?" *Journal of Internet Cataloging*. v. 1, no. 4, 1999, pp. 31-44.
18. Kaplan, Michael, "Exploring Partnerships: What Can Producers and Vendors Provide?" Paper prepared for the Bicentennial Conference on Bibliographic Control for the New Millennium, Library of Congress. November 15-17, 2000. (http://lcweb.loc.gov/catdir/bibcontrol/kaplan_paper.html) (6 Dec 2000) (hereafter cited as "Exploring Partnerships"
19. Some sample comments, alluding to the problem of bibliographic control of titles in aggregator databases:

    "...access to online resources could be improved if all vendors made cataloging records available for the full text products available through their databases."

    "...we catalog journals in publisher-aggregated databases (e.g. Muse, Ideal, etc.) but not in full-text indexes like Lexis-Nexis's various Universes or ProQuest, since coverage, completeness and dates are less than clear or predictable. *IF* we could get solid data, and IF the records of individual titles could be handled "in bulk", it would be great to have those listed in our catalog, but I honestly don't see us being able to add 2000 individual titles (and then keep track of them) for a full-text index."

    "We have an Access database that includes links to more than 19,000 periodicals. We have existing records for the print version of thousands of these titles. It would help us speed things up tremendously if there were a way to do a batch import of the URL's (PURLs) from the Access database to the 856 field, matching on the ISSN for example."
20. Kaplan, Michael, "Exploring Partnerships " (http://lcweb.loc.gov/catdir/bibcontrol/kaplan_paper.html) (6 Dec 2000)
21. For example, the sheer volume of the task; duplication of effort; problems maintaining bibliographic control over e-journals and titles supplied by aggregators. See Barbara Baruth. "Is Your Catalog Big Enough To Handle the Web?" American Libraries. August 2000, pp. 56-60.
22. Baruth, Barbara. "Is Your Catalog Big Enough To Handle the Web?" American Libraries. August 2000, p60.
23. Kaplan, Michael, "Exploring Partnerships" (http://lcweb.loc.gov/catdir/bibcontrol/kaplan_paper.html) (6 Dec 2000)
24. Calhoun, Karen. "Redesign" (http://lcweb.loc.gov/catdir/bibcontrol/calhoun_paper.html) (6 Dec 2000)
25. Collaborative Digital Reference Service. (http://www.loc.gov/rr/digiref/) (6 Dec 2000)
26. Dinberg, Donna. "From CDRS to Document Delivery: a development path toward end-to-end user service," unpublished paper distributed to DigiRef Team 12 October 2000.
27. O'Donnell, Mike, (icopyright.com). "A New Model for Publishing on the Internet," National Online Meeting. Proceedings of the twenty-first National Online Meeting. May 16-18, 2000. p. 303-307.
28. McKiernan, Gerry. "Points of View: Conventional and 'Neo-Conventional' Access and Navigation in Digital Collections," *Journal of Internet Cataloging*, v. 2, no. 1, 1999, pp 23-41.

29. http://www.public.iastate.edu/~CYBERSTACKS (6 Dec 2000)
30. See sources cited at (http://www.public.iastate.edu/~CYBERSTACKS/Agents.htm) (6 Dec 2000).
31. The Scorpion Project. OCLC Office of Research. (http://orc.rsch.oclc.org:6109/) (6 Dec 2000)
32. Information Visualization. Pacific Northwest National Laboratory. (http://www.pnl.gov/infoviz) (6 Dec 2000)
33. MNIS-TextWise Labs. (http://www.textwise.com/) (6 Dec 2000)
34. Kaplan, Michael, "Exploring Partnerships" (http://lcweb.loc.gov/catdir/bibcontrol/kaplan_paper.html) (6 Dec 2000)
35. Cooperative Online Resource Catalog, sponsored by the Online Computer Library Center, Inc. (OCLC). For more information see: (http://purl.oclc.org/corc) (6 Dec 2000).
36. As in the example described by one respondent: "Public services project is identifying which titles in reference collection have electronic counterparts to which catalog needs to link .... we public services/reference librarians need to recommend which links are important to cataloger."
37. Calhoun, Karen. "Redesign" (http://lcweb.loc.gov/catdir/bibcontrol/calhoun_paper.html) (6 Dec 2000)

Library of Congress
January 22, 2001
Comments: lcweb@loc.gov

- Conference Home Page
- What's new
- Greetings from the Director for Cataloging
- Topical discussion groups
- NAS study and 2 articles from the LC staff *Gazette*
- Conference program
- Speakers, commentators, and papers
- Conference sponsors
- Conference discussion list
- Logistical information for conference participants
- Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Caroline Arms

Information Technology Services
Library of Congress
Washington, D.C. 20540

Some Observations on Metadata and Digital Libraries

About the presenter:

Caroline Arms has been at the Library of Congress since 1995, as a technical program coordinator for the National Digital Library Program based in the Information Technology Services division of the Library. In particular, she has been the technical advisor for the *Library of Congress / Ameritech National Digital Library Competition*. Between 1997 and 1999 this competition made awards for twenty-three projects to digitize primary source materials to complement and enrich the Library's American Memory resource. By October 2000, twelve have been integrated into American Memory. Prior to joining the Library, Arms worked at the Falk Library of the Health Sciences at the University of Pittsburgh, as the first Director of Computing at the Amos Tuck School of Business Administration at Dartmouth College, and providing computing support to researchers at the University of Sussex and the Open University (in the United Kingdom). She has a B.A. in Mathematics from Oxford University and an M.B.A. from Dartmouth College. In the late 1980s, Arms edited two volumes for EDUCOM, *Campus Networking Strategies* and *Campus Strategies for Libraries and Electronic Information*, both published by Digital Press.

Full text of paper is available

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

The Internet has stimulated the development and deployment of collections of digital content managed and made available over the network for particular communities or purposes. These digital libraries, with their associated services, have varied ancestry. Some, like American Memory have been built by libraries or other archival institutions. Others have emerged from user communities to provide shared management and networked access for important digital resources, such as survey data for social scientists, sensor data from satellites or telescopes for astrophysicists and other scientists, or instructional resources for faculty and teachers.

The metadata elements needed to allow specialist users to find, identify, select, and obtain the resources they need and to navigate the web of relationships among them do not necessarily match the elements and rules for bibliographic cataloging of materials traditionally held by libraries. The potential for coordinated access to resources of different types from different sources, however, calls for a level of commonality among metadata schemes. Simple and rapid access to full content may reduce the need for some cataloging details, since the user may be able to use the full content or an automatically created summary, such as a thumbnail of an image or outline derived from marked-up text, to aid selection. On the other hand, although archival collections in paper form are often described as a whole or at the level of a series or physical container, item-level identification is essential in a digital library, increasing the cataloging cost. However, content in digital form can be a source for automatically generated metadata; such metadata will be less costly but flaws that would be easily corrected or avoided by a human cataloger may go undetected. In digital libraries, not all relationships between items have to be recorded in catalog records. Relationships between digital works can be embedded when the work is created or derived automatically by analysis of the full content. Citations can link to the works referenced, providing navigation capabilities far richer than those possible through catalog records.

This paper will draw on experience gathering together metadata from heterogeneous sources for American Memory, particularly for the collections digitized and cataloged at other institutions through the LC/Ameritech competition. It will also reflect on several initiatives to develop rich structured metadata schemes for specific domains and others to find simple approaches to support resource discovery across domains. Trends and commonalities will be identified and influences among metadata schemes highlighted.

Library of Congress
June 27, 2000
Comments: lcweb@loc.gov

Some Observations on Metadata and Digital Libraries

Caroline R. Arms
Information Technology Services
Library of Congress
Washington, D.C. 20540

Final version

> "By the year 2000, information and knowledge may be as important as mobility. We are assuming that the average man of that year may make a capital investment in an "intermedium" or "console"--his intellectual Ford or Cadillac--comparable to the investment he makes now in an automobile, or that he will rent one from a public utility that handles information processing as Consolidated Edison handles electric power. In business, government, and education, the concept of "desk" may be primarily a display-and-control system in a telecommunication-telecomputation system--and its most vital part may be the cable ("umbilical cord") that connects it, via a wall socket, into the procognitive utility net."

J. C. R. Licklider. *Libraries of the Future*. M.I.T. Press, 1965

The words above may not be those in use today, but the prescience of these sentences from thirty-five years ago is amazing. At the time, computers were used by "collecting data and writing a computer program, having the data and program punched into cards, delivering the cards to a computer center in the morning, and picking up a pile of 'printouts' in the afternoon." Time-sharing computing systems with typewriting terminals for remote users were just emerging from the research laboratory for practical use, as was the use of cathode-ray devices as terminals. The book, *Libraries of the Future*, was based on a two-year study sponsored by the Council on Library Resources and carried out, under Licklider's leadership, by a group of engineers and psychologists from Bolt Beranek and Newman, Inc. (BBN) and the Massachusetts Institute of Technology (MIT) starting in late 1961. Charged by the Council to explore how developing technologies might shape libraries in the year 2000, the group envisioned a much closer "interaction with the fund of knowledge"(1) than print libraries can support. They saw the fund of scientific knowledge directly available not only to scientists but to their experiments; they envisioned the ability to feed research results directly back into the fund of knowledge. Licklider would surely be delighted to see the systems for accumulating and using genome resources today. This vision was so different from the library of the early 1960s that it seemed appropriate to use a different term; the term chosen was "procognitive system." In the year 2000, the Internet, with all the information resources to which it provides access, serves as the "procognitive utility net" that Licklider predicted.(2) Communities, institutions, and individuals have been building digital libraries as they work towards the vision they share with Licklider and his colleagues for a richer, closer interaction with the fund of knowledge.

Libraries have always supported interactions with the fund of knowledge, interactions that come in many shapes and sizes. Libraries support scholarly communication and formal education; they also help people find facts, figures, and tax regulations. Interacting with knowledge is what lifelong learning is all about. Users of American Memory register delight when they find pictures of the town where they grew up or recognize a family member in a picture, sound recording, or letter. Genealogical research is immensely popular. Contributions to the fund of knowledge come from many sources, including individuals as amateurs. Vast numbers of informative web pages represent personal contributions to the fund of knowledge by enthusiasts: railroad buffs; music-lovers; naval history mavens; watchers of birds, badgers, and other creatures; and many more.(3) I will not argue that the World Wide Web is a digital library (although some do). I will limit my concept of a digital library to collections of resources in digital form assembled for a particular community or purpose and managed with an intention of ongoing accessibility and usability. With this loose definition, perhaps what distinguishes a digital library from a set of documents or web pages is the existence of some formalized, structured metadata (data about data) to provide organized access to a body of resources.

User communities are building domain-specific digital libraries with domain-specific metadata schemas and guidelines.(4) This should be no surprise. Domain-specific controlled vocabularies and abstracting and indexing services are not a new phenomenon and the full-text databases that some of these services have developed into are certainly digital libraries. Even within the traditional cataloging community, descriptive practices vary for classes of material. Practices developed originally for cataloging books have been adapted and extended over 150 years (if one considers the plans developed by Sir Anthony Panizzi for organizing books in the British Library as the starting-point for modern bibliographic practice). As Elaine Svenonius points out in her recent book, *The Intellectual Foundation of Information Organization*, these practices "have been jolted in the twentieth century by information explosions, the computer revolution, the proliferation of new media, and the drive toward universal bibliographic control. How they have withstood these jolts, where they have remained firm, where they have cracked, and where cracked how they have been repaired or still await repair is a dramatic -- and instructive -- history for those interested in organizing information intelligently." (5) For example, in response to the jolt of incorporating non-book media, the Anglo-American Cataloguing Rules have been supplemented by manuals for other classes of material: *Archives, Personal Papers, and Manuscripts* (APPM); *Graphic Materials* (GIHC); and *Archival Moving Image Materials* (AMIM). An archival collection of personal papers is typically cataloged in a single collection-level record. Very different descriptive and organizational practices have been developed by archivists to organize and describe collections at whatever finer level of granularity is deemed appropriate.

Varying descriptive practices have taken into account not only the observed intellectual needs of the traditional users of the resources, but also more pragmatic factors: the mission and capabilities of the traditional custodian; economic realities (manpower and funding); technical realities (tools available to help custodial institutions prepare and users to take advantage of metadata); the pattern of updating required; and the physical nature of the artifacts themselves. The nature of today's bibliographic systems make allowance for synergies with (or are constrained by exigencies of) inventory control and are shaped by the importance in the published literature of relationships expressed by shared author, title, and subject headings. For archivists, two of the key factors shaping practice are the importance of the integrity of a collection as a whole and the sheer impracticality of describing individual items in detail. Although catalogs in book form were replaced by card catalogs early in the twentieth century, the document form of an archival finding aid has remained useful. For museums, the importance of detailed information about provenance and the historical and creative context for individual items has led to yet another set of practices. Many of these factors do not change simply because reproductions or surrogates can be created in digital form or when today's analogs are created in digital form. Even among the communities with preservation of the cultural record in their mission, there will continue to be heterogeneity of descriptive practice in digital libraries, and for good reason. One challenge is to identify the pragmatic factors that have changed or will inevitably change and adapt to them. Underlying principles grounded in users' needs should still guide practice.

What is different in a digital library?

The networked world of cyberspace and the development of advanced computational tools are shifting the balance among factors that shaped past descriptive practices. Although developed for a different purpose, the model Lawrence Lessig introduced in *Code and other Laws of Cyberspace* sheds light on this balance (for this author, at least). Lessig suggests that an individual's behavior is constrained by *law, architecture, market, and norms*.(6) He stresses the fundamental differences between the architecture of physical space (where distance, weight, and walls set limits) and the architecture of cyberspace (manifested in software, network equipment, and protocols). The four factors are clearly interdependent; indeed, in response to societal norms, laws are passed to regulate architecture and the market. Turn to the business section of any newspaper, and it becomes obvious that the architecture of cyberspace is changing the market (and the overall economic environment) in which businesses, individuals, and libraries operate. Cyberspace offers new cost structures, users from new communities, users from traditional communities with new expectations, and new tools to serve those users. Clearly, understanding of the surrounding "architecture" and "market" will guide the development of future practices and systems for preparing and using metadata in digital libraries.

The architecture and market for print publishing has been relatively stable and libraries are adapted for that environment. For digital libraries, the environment is likely to be in a state of change for the foreseeable future; research and experimentation will be ongoing. For a thoughtful analysis of some of the metadata issues warranting research, see *Metadata for Digital Libraries: a Research Agenda*, developed by a joint task working group established under the auspices of the European Union and the National Science Foundation.(7) The current article should not be seen as an attempt to develop an overarching theoretical or technical framework or as a comprehensive overview, but as observations from the trenches of American Memory, a production digital library system that is also an experiment. The integration of heterogeneous content, including content and metadata prepared by other institutions, into American Memory has provided a close look at practical hurdles in the path to Licklider's vision. It has also stimulated an appreciation for how varying descriptive traditions can contribute to achieving that vision and a not infrequent sense of frustration at the difficulty of incorporating new tools to build better services and enrich the interaction with the fund of knowledge.

Objectives for metadata and expectations of users

In the IFLA *Functional Requirements for the Bibliographic Record* four objectives are listed for the bibliographic record for an entity: to enable a user to *find, identify, select,* and *obtain access* to the entity described. Elaine Svenonius suggests that it is helpful to distinguish between the objective of *finding* or *locating* a particular entity (known item) and the *collocation* objective, which allows a user to find sets of entities, for instance works by the same author or on the same topic.(8) She also suggests that a *navigation* objective be added to reflect the wish of users to find other works related to a given work.(9) The Dublin Core Metadata Initiative(10) describes the Dublin Core metadata set as facilitating *discovery*; this term is often applied to the process of looking for resources in a broad cross-domain universe. In a digital library, metadata may be needed to help the user not only *select* an item appropriate for the current purpose but also to *use* it. Such metadata may include structural metadata that permits navigation among components of a single intellectual "object" (for example, to skip to a particular segment of a digital sound recording) or information about terms and conditions in a machine-parsable form that can be used to limit access to authorized users. Metadata is also required to manage content and to record information that will support future preservation activities. In this article, the focus is on metadata that leads users to resources in digital libraries (*discovery, finding, collocation, navigation*) and lets them choose resources for the task at hand (*identification, selection*).

The networked world has changed the expectations of information users by removing physical constraints. In the physical world, few question the practice of storing maps or manuscripts separately from books or locating the specialists who help users access and use the resources in different areas of a library or in different libraries. In the networked world, the digital libraries hold the potential to brings resources of many types together to the user's laboratory or desktop (or even palmtop). Some users will want first to cast a wide net as they trawl for information and then to draw it tight and precisely around the most relevant and usable resources for a particular purpose. The wide net calls for interoperability and commonality across domains and custodial communities, while assessment of fitness for a particular purpose often calls for rich and domain-specific metadata.

One characteristic of a digital library is the accessibility of the content to the user. In the traditional library, subject to the physical constraints of weight and distance, the user who selects lots of items that turn out not to be fit for the purpose at hand will incur a high cost (in effort, if not in monetary terms). If users can delve straight into the content to aid the act of selection, some metadata conventionally recorded to support selection may be less essential. Users and builders of digital libraries also recognize the potential for navigation among related items. This feature was the fundamental function underlying the initial success of the World Wide Web. Online, citations can lead directly to cited works and works can link directly to datasets or models they describe. Reference links and thumbnail images change the architecture for research and information seeking. Users can navigate using relationships between metadata records and relationships between resources from their desktop. The functions that metadata must support may not change in a digital library, but, in the new architecture and with new tools, it seems likely that metadata practices must evolve to provide the functionality cost-effectively in the new market.

Content is also accessible for building services in a digital library. Metadata or surrogates can be derived or extracted from content files. Today, a thumbnail is essentially part of the descriptive record for an image. In the collection of negative strips from the Farm Security Administration, many of the images were never captioned; access to the uncaptioned images within American Memory is primarily through navigation. Contact sheet displays of thumbnails are ordered using the processing number sequence on the negatives (an identifier that implies chronological order, at least for negatives on a single strip). Automatic summarization and analysis of other forms of content is an active research area. The value of including basic metadata in the header of text marked up in SGML or XML is also well recognized. Text conversion projects following the Text Encoding Initiative (TEI) guidelines often use a mapping between elements in the header section of the marked up file to fields in a MARC catalog record. In some cases, the TEI header is derived from the catalog record; in other cases, a catalog record is derived from the header. For material published in digital form, the publishing community will maintain basic bibliographic metadata for its own purposes, including promotion and business-to-business dealings with booksellers. It will be inexpensive to include it in file headers. For example, the Open eBook Publication Structure specification includes tags for "publication metadata."(11) In a digital library, searchable text can minimize the need for metadata (particularly when items do not merit the expense of individual cataloging). American Memory has a popular collection, *American Life Histories*, for which the only item-specific metadata is a title (a display string needed for result lists) and an identifier; the text itself is searched to provide intellectual access.

Creators of non-text materials also recognize the value in embedding such metadata within the files. Today, cameras can record the date and time of an exposure. If there is not already a camera with built-in GPS to record location, there will be very soon. Proposals for new digital file formats, such as JPEG2000 (12) (for images) and MPEG-7 (13) (primarily for sound and video) include the ability to embed descriptive metadata within the file.

Community-specific metadata models and schemas

The educational community has been active in exploring the functional needs of educators for finding and using instructional materials. Resources are available because of economic incentives to establish and manage online learning environments and government efforts to improve education. Some information tasks call for rich metadata. For example, a teacher may be looking for material to help him explain a very specific topic (say cell division in an embryo) in a particular class. The teacher wants to be confident that the material makes appropriate assumptions about what the students already know, has already been used successfully in the classroom, will occupy an appropriate amount of class time, and will work on the equipment available. This level of specificity may be available in a service built by and for educators. A metadata schema for instructional resources has been developed by the IMS Global Learning Consortium, Inc., a global consortium with members from educational, commercial, and government organizations.(14)

Another area with specialized metadata needs is that of geospatial resources. The ability to build maps dynamically or relate geospatial facts from distributed sources of information is enhanced by commonality of metadata. Such capabilities can support government tasks such as emergency management, city planning, and tracking air quality or global climate change. The Federal Geographic Data Committee has developed a Content Standard for Digital Geospatial Metadata.(15) There is worldwide standardization activity in this area in relation to metadata schemas and to the content standards for elements. The activity involves government, commercial and educational sectors. As with the instructional metadata schemas, the functionality required by the creator and user communities, rather than traditional libraries, is driving these activities.

The Visual Resources Association has developed a two-level hierarchical model for describing objects or visual works (such as paintings, sculptures, or buildings) and images of those works. A single set of metadata elements, the VRA Core Categories 3.0 (16), can be applied to the works and to the images. This approach follows the so-called "1-1" principle, distinguishing characteristics of the image surrogate clearly from characteristics of the work. This principle emerged from the Dublin Core community but has provoked considerable controversy. It can not be applied in existing "flat" bibliographic systems without creating awkwardness and confusion for users through multiplicity of records and a burden on cataloging processes by requiring replication of work-level information in records for each surrogate image. A system that takes advantage of the two-level structure must allow searching across all records but present results that pull in work records automatically when a "hit" is at the image level. The metadata schema used by the Art Museum Image Consortium (AMICO) has a similar hierarchical structure.(17) The Visual Information Access (VIA) system at Harvard University uses a three-level hierarchy.(18) Although the Encoded Archival Description (EAD) standard has been used primarily for describing collections of papers and records that have not been digitized, it too provides a hierarchical structure for description at different levels.(19) The EAD metadata structure has been used effectively as the basis for digital library services, for example at the University of California, Berkeley(20) (and now at the California Digital Library) and Duke University.(21)

A more complex conceptual model has been proposed by the Documentation Standards Group International Committee for Documentation of the International Council of Museums (ICOM-CIDOC).(22) The CIDOC model is an object-oriented reference model that expresses a much more complex knowledge universe than the simple relationship of a descriptive record to a resource or even of a hierarchical structure of related descriptions and resources. It allows for descriptions not only of works, images, document, (conceptual objects) and objects (physical entities), but also of people (actors), places, and periods (time-spans). This model will form the basis for the Cultural Materials Initiative digital library project at the Research Libraries Group. Full use of such a model will integrate gazetteers, biographical dictionaries, and encyclopedias, going well beyond the traditional use of authority records and thesauri.

One example of a digital library enriched by integrating the use of reference resources such as biographical dictionaries and the Thesaurus of Geographic Names is *Perseus*, housed at Tufts University.(23) In this digital library for the study of the ancient world, the books "talk" to each other. Names of places and people been identified automatically in the text and can be used as links to related information elsewhere in the corpus. This includes automated disambiguation of different places with the same name (e.g., Springfield) and of references to people who share names with places (e.g., Lincoln). Licklider would have been delighted. In 1965, he complained that "when it comes to organizing the body of knowledge, or even indexing and abstracting it, books by themselves make no active contribution at all."(24)

Some of these rich metadata schemas or models have great potential for enriching the interaction with knowledge for users through collocation and navigation using the relationships expressed in the models and supporting knowledge organization systems, such as gazetteers, name authority files, and thesauri. The models are, however, unfamiliar and will require new tools to implement and deploy. The web sites that present them usually have FAQ (Frequently Asked Question) pages that emphasize that most features are optional. There is plenty of evidence that, unless there is strong economic motivation, introductory guidelines and basic tools are needed before complex standards gain broad acceptance. Part of the genius shown by Tim Berners-Lee in his original standards for the World Wide Web was the simplicity of the specifications. The conceptual model could be explained on one slide and fleshed out in three more. Implementing a server was a few lines of code; a young programmer built a graphical interface (Mosaic) as a side project and the rest is history. The other aspect of Berners-Lee's genius was the instant integration of legacy content by supporting earlier Internet protocols now seldom mentioned (gopher, wais, news). Real-world experience with the simple (e.g. HTML) has led iteratively to better understanding of which extensions to its functionality are most essential. Complex metadata schemas present a challenge for those with valuable legacy metadata to migrate or metadata maintained in rich, but different schemas. It is relatively easy to "dumb down" metadata records from a rich schema into a simpler one, for purposes of interoperable retrieval, while still maintaining the full richness in a master system. Transforming from one rich schema to another is usually more expensive and the benefits may not be obvious to the institutions or individuals most likely to bear the cost.

Metadata for cross-domain discovery

At the other end of the spectrum, some digital library activities are focusing on allowing users to find resources across an information universe that spans communities, nations, types of information, and types of institution. The Dublin Core Metadata Initiative has been building consensus through a series of workshops and working group activities since March 1995. From the start this has been an international effort to develop a common core of semantics for resource description. The path has not been easy. Active debates have highlighted the differing priorities and expectations of different communities. The Dublin Core Element Set (Simple Dublin Core) was submitted to NISO as a draft standard (Z39.85-xxx) this year.(25) The set has 15 elements (listed in the left-hand column in Table 1); all elements are optional and all repeatable. The elements themselves are unlikely to be sufficient as an internal metadata schema for any particular project or application. Some proponents see the elements as the basis for a schema that can be extended by adding or refining elements; others prefer to see the set as a view of a richer, more complex description.(26) This view can be used as a framework for mapping different element sets into a common set for indexing and searching. The DCMI has also developed a model for extending the simple element set while maintaining the objective of interoperability. Elements can be refined. For example, Date.Created and Date.Issued are refinements of Date. Element refinements are guided by the "dumb-down" principle. If the refinement term is not recognized, it should be reasonable to treat the value of the qualified element as if it had no qualifier. This is an extremely important principle for supporting broad interoperability. The second type of qualification for Dublin Core elements is to specify an encoding scheme or controlled vocabulary. In July 2000, a set of exemplary qualifiers was published as a result of proposals made by working groups at the DC-7 workshop in October 1999.(27) Tom Baker describes Dublin Core as a pidgin language in a discussion of simple and qualified Dublin Core.(28) The existence of the 15-element set has provided an important focus for other interoperability initiatives.

Simple Dublin Core was recently adopted as the core metadata record format for the experimental Open Archives initiative. This initiative is testing whether a simple mechanism that allows service providers to harvest metadata records from content repositories will facilitate and stimulate the development of valuable services that draw on content from many repositories. An initial impetus was the belief that access services (such as reference-linking, portals, and selective dissemination services) could benefit from harvesting records for e-prints and other "grey" literature. Records will usually have links back to the host repository through persistent identifiers for the full content. The concept also allows the development of comprehensive search services that include significant web-accessible resources currently hidden from the "spiders" that crawl the web on behalf of search engines. The so-called "deep web" includes the content of most digital libraries, such as American Memory, whose web presence is largely ephemeral, with records retrieved from a database in response to each search and displays generated dynamically. The Open Archives harvesting framework includes the ability to harvest records in other metadata formats (e.g. MARC or the rfc1807 bibliographic format used for the Networked Computer Science Technical Reports Library). The Open Archives initiative is based on the premise that simple records (almost certainly derived dynamically from a more complex schema used internally) provide the first step to cross-domain discovery. Service-builders can choose to harvest records in a richer schema when available. Specifications for records marked up in the Extensible Markup Language (XML) must be made accessible for any schema used.

A third activity aimed at cross-domain discovery is the development of a very basic profile for the Z39.50 information retrieval protocol for cross-domain discovery. This specification also supports Simple Dublin Core marked up in XML as a transfer format for records. In many ways, the stimulus for this activity is the success of the World Wide Web. People clearly find value in web search engines, whatever their shortcomings. An enormous mass of information accessible from a single search box has clear appeal; many people prefer to try several queries and skim through pages of hits than to use Boolean queries to increase precision. The computational and linguistic tools built into commercial search engines are enhanced frequently. The architecture of cyberspace has changed the relationship between bibliographic control and access. Today, ironically, resources under good bibliographic control are likely to be less widely accessible than those simply mounted on the web. The motivation behind these simple-minded interoperability efforts is to encourage broad access to resources of value.

Types, formats, and genres of digital content

The fund of knowledge is represented by a much richer set of resources than static pages on paper, and resources beyond those traditionally found in libraries, even multi-faceted libraries like the Library of Congress. Knowledge has always been conveyed through buildings (and the archaeological sites they become), works of art, physical specimens of flora and fauna, artifacts of different cultures and lifestyles, and human memories. Photography, sound recordings, and motion pictures have added to the fund of knowledge both in their own right as means of expression and as richer surrogates than words on paper. The digital era has added not only reproductions and analogs of older forms of information, but also new digital resources of enormous variety. Some fall into obvious categories. The broad category of datasets includes census and other survey results, gene sequences, images and other sensor data from space, geospatial information that allows maps to be generated dynamically, decades of financial statements for publicly traded corporations, directories of people and places, and structured lexical resources, such as dictionaries and thesauri. More complex digital resources include mathematical and chemical knowledge expressed in structural forms that permit dynamic manipulation (and the software that performs or lets users perform the manipulation), interactive software for education and entertainment, and collections of re-usable "open source" software code. Digital libraries are being built to manage, serve, and support discovery of all these categories of resource. Some of these digital libraries are extensions of traditional libraries; many have developed from other well-established activities in organizing information (for example, for collections of social science datasets). Dynamic information resources, such as the web site that delivers up-to-the minute details of event and results at the Olympic Games challenge all traditional practices for organizing and recording for posterity. However, this too could be represented as a series of snapshots of bit-patterns, a digital resource, a set of computer files. The metadata required to support discovery and use of digital resources must clearly represent the intellectual nature and genre of the content; in a digital library, the fact that a resource can be represented by 0s and 1s is an assumption, not a useful categorization.

Attempts to develop general hierarchical categorizations for genres or types of information have usually failed. Even within American Memory, the content does not fit into a neat hierarchy. Are maps a subclass of images? How do you relate page-images of sheet music, song transcriptions, and recordings of performances? The Type working group of the Dublin Core Metadata Initiative developed, with much debate and without unanimity even in a small group, a high-level list of types (the DCMI Type Vocabulary: DCT1): Collection, Dataset, Event, Image, Interactive Resource, Service, Software, Sound, Text.

MARC records can hold information about the type of a resource in several ways. Svenonius notes that there are seven different places where a "document type" can be indicated.(29) Each element or indicator has a different set of possible values. Given the different guidelines for use and the different functions these seven elements serve, type information in MARC records has proved impossible to use uniformly within American Memory. Type indicators in the header are excellent as triggers for systems based on MARC records. They indicate what guidelines have been applied to content fields and therefore can be used to configure appropriate displays or procedures. However, the coded values can not be incorporated into a general keyword index, and are therefore unavailable to a user as a search term as they expect. Elsewhere, Svenonius remarks, "the use of one device to serve multiple functions, ..., while favored by the principle of parsimony, nevertheless introduces a lack of flexibility that can be an obstacle as technology changes."(30) The principle of parsimony results in type information for some classes of material appearing only within a complex physical description not designed for machine parsing.

The DCT1 list by itself would not prove sufficient for item-level genre distinctions in American Memory. Several of the categories don't apply to a body of converted analog materials and those that do not support the *selection* objective adequately for the typical American Memory user. Svenonius argues, "For document types, as for general-format types, it is not possible to construct a classification that is both natural and whose categories are mutually exclusive." American Memory experience supports her argument. The current feeling of the Dublin Core Type working group is that some communities will develop controlled lists of terms, but agreement across communities on the important categories or even common definitions for the same terms is unlikely. Lacking initial agreement on an acceptable "standard" typology, American Memory does not have explicit type values in all metadata records. This shortcoming means that searches limited to images may retrieve other categories of content, since the limit is actually by collection and some collections included will have text or sound as well as images. Colleagues agree that adding high-level type information consistently across the metadata records would provide more benefit for American Memory users than any other change. Interoperability will certainly be well served if descriptive records shared or exchanged always include type information, even if all content within the repository or collection providing the data is of the same type. Based on experience with American Memory, users might be best served by the inclusion of any applicable terms from the DCT1 list **and** additional terms at finer levels of specificity.

Metadata for search and metadata for display

The user's objectives are supported not by raw metadata, but through the tools and systems that can take advantage of the metadata. In both library catalogs and digital library services, the functionality for discovery, finding, and collocating is determined not by the metadata but also by the indexes constructed for that metadata. Different systems provide different options for configuring indexes. Public interfaces to library catalogs usually combine different metadata elements (e.g., MARC fields and subfields) into a relatively small set of indexes. For example, a keyword search by subject may find the term in any of the MARC subject fields (e.g. personal names, terms from authorized vocabularies, uncontrolled terms, genre terms, etc.). Once a record has been retrieved, elements can be labeled more specifically.

Some digital libraries, including American Memory, take the same approach. Individual elements that may be usefully distinguished (and labeled) to support the act of selection are lumped together to support discovery, finding or collocation. The University of Washington Libraries have used collection-specific element sets for their collections of digital reproductions; each set is mapped to the fifteen elements of the Dublin Core Metadata Element Set for cross-collection retrieval.(31) Retrieved records show the specific labels. For example, a collection of pictures of plants has a variety of fields relating to preferred soil quality and climate, whether the plant is native to the state, and other botanical details. For search purposes, these fields are all included in the Description index; on the display they are individually labeled. American Memory uses a similar approach. All descriptive notes are lumped into an overall text index; on display, a summary or abstract is usefully presented first and labeled as such. Many collections in American Memory call for unusual metadata elements, such as musical features for folk songs and descriptions of the key mechanisms in a collection of flutes. For searching, these are all treated as notes and included in the general keyword index. For display, however, the metadata format includes tags that are ignored by the indexer, but provide labels for use in the record display.

The indexing approach currently used in American Memory was developed empirically and iteratively, based on data elements usually available across heterogeneous sources and expectations of content custodians and users; it has seven primary indexes that support searching of metadata (Title, Alternative Title, Creator, Contributor, Subject, Any Text, Number). The developers of the Alexandria Digital Earth Prototype at the University of California, Santa Barbara have described the framework they use for querying metadata from distributed sources. Based on experience over several years of working with geo-referenced information, they chose eight search buckets (Geographic Location, Type, Format, Topical Text, Assigned Terms, Originator, Date Range, Identifier). For the most basic cross-domain discovery, the developers of the Bath Profile for Z39.50 identified Author, Title, Subject, Any. The Bath Profile effort is not strictly a digital library project, but has as an aim, interoperability between library catalog systems and "other electronic resource discovery services." Table 1 provides an informal tabular comparison of the clusters for indexing of these three projects and the alignment with Dublin Core.

Table 1: Comparison of search buckets for metadata for digital library projects

Dublin Core Metadata Element Set	**American Memory (local search buckets in parentheses)**	**Alexandria Digital Library (search buckets in bold)**	**Bath Profile (Z39.50) for Cross-Domain Discovery**
	Digital library of reproductions of historical sources (in text, image, sound, video)	Distributed digital library for geographically-referenced information	
Elements that support discovery			
Title	**Title** (TITL) **Alternative Title** (ALTTITL), usually searched with TITL.	(Topical text)	**Title**
Creator	**Creator** (AUTHOR)	(Originator)	**Author**
Contributor	**Contributor** (OTHER), usually searched with AUTHOR	(Originator)	
Publisher	(TEXT)	(Originator)	

Date	Display date (TEXT) Sort date (used only to sort search results within collections for which dates are known well enough to normalize)	(Date range)	
Subject	**Subject** (SUBJ)	**Assigned terms**	**Subject**
Coverage (spatial and temporal)	Geographic subject. (hierarchical placename, indexed as SUBJ, also used to support browsing of placenames and map-based selection by state)	**Geographic location** (footprint in geographic coordinates) **Temporal coverage** (indexed as Date range)	
Description	Summary (TEXT). Other notes (TEXT)	(Topical text)	
Type (genre)	(SUBJ)	**Type**	
Language (of resource)	Language (TEXT)		
Format (digital)		**Format** (for delivery, online or offline)	
	Any text elements, including textual fields in other indexes. (TEXT)	**Topical text** (includes title, description and any other text, including assigned terms)	**Any**
		Originator (includes creator, contributor, publisher)	
		Date range (includes date and temporal coverage)	
Elements that primarily support identification and navigation and use			
Identifier	**Identifier** (NUMBER)	**Identifier**	
Relation	Related items		

Source			
Rights			
	Repository	Reproduction number (NUMBER)	

How do users search in digital libraries?

The columns in Table 1 are compromises, reflecting a balance between what users ask for and what the metadata can support. Reading between the lines, I see a much stronger similarity in desired functionality for American Memory and the Alexandria Digital Library than the table would imply. The only significant difference is that American Memory users do search for titles, for example, for books and songs. With that exception, the search buckets used for the Alexandria Digital Library (ADL)(32) would suit American Memory users well -- if the metadata were more consistent. [The separate index for alternative titles in American Memory is for efficiency; it allows the search engine to generate hit lists based on titles without retrieving the full records.]

Searching by topic (assigned terms and almost any text)

The common text bucket proves invaluable when dealing with heterogeneous data and it is useful to include subject terms in this bucket. Indexing engines designed for full text will find word variants automatically, relieving users from knowing when formal subject terms use the plural form. The distinction between assigned subject terms and textual description is, nevertheless, valuable. In American Memory, it allows us to generate browse lists, which are actually static (but easily and automatically regenerated) HTML pages with "canned" searches. The subject index also permits navigation from subject terms on record displays to other records assigned the same terms.

Searching by originator

Although American Memory indexes the primary Creator separately from the other Contributors, searching and browsing for creators and contributors is usually done together. Combining the indexes has been considered. In American Memory, the addition of roles to Creators and Contributors proves valuable, particularly for non-text materials (e.g. to distinguish composer from lyricist or illustrator of sheet music). Authorized forms for names are extremely valuable in American Memory. However, names within text are also an important access point.

Searching by date range

The Alexandria Digital Library is designed for powerful searching by geographic location and date range. Strict machine-parsable content standards are used for those metadata elements. American Memory users would love to be able to search more effectively by date and place. For much content in American

Memory, unfortunately, dates of creation for the original item are uncertain and date-ranges recorded are often so broad as to be useless for discovery. For certain collections where chronology is important, normalized dates have been generated and can be used to sort search results. It is interesting that the Alexandria Digital Library has chosen to index date of creation/publication in the same bucket as date of coverage. In American Memory, there are many instances in which the dates are essentially equivalent (for example, the date a photograph was taken, a letter written, or the proceedings of the Congress recorded). For published books and maps, the distinction is often important, but users may be interested in either or both. The Alexandria Digital Library uses special overlap and containment queries for date ranges and location. Such queries would not be efficient with a full-text engine, but American Memory users would like the capability.

Searching by place

Geographical location is an area in which traditional descriptive practices aimed at human-readable displays do not transfer well to digital libraries (or even support finding and collocation in traditional library catalogs). In American Memory's metadata from heterogeneous sources, location may be expressed in many ways: as an informal place-name, as a subdivision in a topical subject heading, as a traditional subject heading (e,g., Brooklyn (New York, N.Y.)) or a hierarchical place name (e.g., country -- state -- county --city). Of these, by far the most useful for manipulating automatically with simple text-based tools has been the hierarchical place name. We look forward to being able to use gazetteers to convert place-names to bounding boxes (coordinates) of the type used by the Alexandria Digital Library.

Searching or limiting searches by type

Users often know that they are looking for an image or for a map and would like to exclude other types of information from the start. Searching by more specific genre terms is also useful, for example for posters or cartoons. As pointed out earlier, type categories expressed in terms that users recognize should be available for limiting or searching. Controlled vocabularies are useful, but are likely to be domain-specific. Convenient distributed access to vocabulary or authority registries is a part of Licklider's vision that has not yet been achieved.

Improved tools to support access to resources in digital libraries

Access will be enhanced through better tools for generating and transforming metadata, better tools for sharing and exchanging metadata, better tools for search and retrieval, and better tools for post-processing search results. The emergence of XML (Extended Markup Language) and its widespread support as a syntax for exchanging metadata and content, particularly for e-commerce transactions and services, is stimulating the development of better tools for transforming and sharing metadata. This, in turn is leading to support for XML from vendors of database software and text-indexing engines.

It appears clear that XML will provide the syntax for metadata exchange among digital libraries in the coming years. A few examples of its adoption to support interoperability include: the Bath Profile (as a record syntax option for search and retrieval using Z39.50); MPEG-7 (for the MPEG-7 Description Definition Language); the Open Archives Initiative (to allow information service providers to harvest metadata from data providers); and the Federal Geographic Data Committee(33) (as the syntax for the Content Standard for Digital Geospatial Metadata). XML is also being used as a syntax to represent content objects, such as the proposed Open eBook standard. American Memory has relied heavily on common indexing of heterogeneous metadata, with different element sets and in two different digital formats (MARC communications format and an XML-like syntax for simple (Dublin Core-like) records. Migration to an XML syntax with a formal DTD or schema is anticipated.

Transformations from one XML metadata schema to another (especially from a rich one to a simpler one) can be facilitated using Extensible Stylesheet Language Transformations (XSL/T). This is just one example of the powerful XML-based tools emerging. XSL/T is already in common use for transforming finding aids marked up to the EAD standard (in XML) to HTML for display on the web. XML also includes the ability to mix and match metadata element definitions (semantics and syntax) from other schemas (using the "namespace" feature). As the flurry of XML-related activity on the World Wide Web Consortium web site in late 2000 shows, the general acceptance of XML signals the beginning of a further period of experimentation and development. Among the unanswered questions relating to the use of XML as a syntax for metadata is how soon (or whether) there will be widespread adoption of the Resource Description Framework, an elegant modeling framework for descriptive schemas.(34) RDF is layered on top of XML, using a particular XML-based syntax for metadata. RDF-specific tools will be needed to take full advantage of its potential for scalable interoperability. Whether the mix-and-match potential of XML namespaces will be widely exploited also remains to be seen.

In the past, text-indexing engines, relational database management systems, and SGML-based storage and retrieval systems all offered different functionality to support the finding, collocation, and discovery objectives for digital libraries. A text-indexing engine can handle heterogeneous metadata and full text in a single system; the capabilities for matching word variants (stemming) have been invaluable for American Memory. Features standard in relational database systems, however, such as sorting by date, have been implemented by additional programming. SGML-aware systems have advantages for substantial bodies of highly structured textual content, such as books and periodicals. Recently, products in one category have found ways to integrate the capabilities of another. ORACLE now has a full text search module, CONTEXT. Some text-indexing engines can now index text stored within relational databases. All such products are announcing "support" for XML.

The other area where tools are emerging is in automated integration of thesauri and other knowledge bases to support more intelligent retrieval. Such tools can compensate for less complete metadata. These knowledge bases could also be more widely used as a resource when creating metadata. Digital libraries will benefit from network-accessible thesauri and authority files that can be queried dynamically from systems that are used to generate metadata (whether automatically or by human catalogers).

Looking ahead

The vision that Licklider and his colleagues expressed in 1965 of libraries that allowed richer "interaction with the fund of knowledge" is still a goal to strive for. I make no attempt to look ahead another thirty-five years. From the trenches, my view toward the horizon includes rich metadata schemas for content that warrants it, simpler schemas that encourage broader access to organized knowledge through interoperability, and ongoing popularity of simple-minded searches supported by intelligent tools in the background. As communities develop rich metadata schemas, I hope they take advantage of the existing fund of knowledge on organizing knowledge. Elaine Svenonius looks back on the history of cataloging and discusses principles, practices, and most refreshingly, problems with practices. Picking up on one of the problems she mentions, the shortsightedness of using one "device" to serve multiple purposes, I offer a few snippets of advice to those designing or applying metadata schemas. I present the advice in my own words, but am confident that many of my colleagues in the National Digital Library Program share the sentiment, because it reflects frustrations they have expressed.

In metadata schemas, draw clear distinctions between elements that serve different purposes. Some examples from the American Memory experience include:

Content type (genre, mode of expression)	Types used to identify a set of guidelines used for cataloging or description.
	Types used to trigger behavior in a particular system or application.
	Types as terms for users to use in queries.
Dates and periods	Dates or date ranges intended for automatic manipulation, such as sorting or access through a timeline slider.
	Dates, date ranges, or periods intended to be readable by humans.
Geographic locations	Coordinate-based locations, that can be used (a) in a map-based query interface that retrieves items ranked by distance from a query point or (b) to respond to queries that look for inclusion within or overlap with geospatial footprints.
	Hierarchically structured names that can be used for simple map-based querying and for conversion to geospatial footprints using gazetteers.
	Place names intended to be read by human users.

Finally, I would like to express my thanks to the organizers of the conference on *Bibliographic Control for the New Millennium* for asking me to contribute a discussion paper. Without this stimulus, I might never have read *The Intellectual Foundation of Information Organization* by Elaine Svenonius from cover to cover. Her deep experience and shrewd analysis shed light on our struggles with heterogeneous metadata in building American Memory and provide articulate confirmation for some of the lessons we have learned from experience. This is what "interacting with the fund of knowledge" is all about.

References

Baker, Thomas. (2000). "A Grammar of Dublin Core." D-Lib Magazine, October 2000. [http://www.dlib.org/dlib/october00/baker/10baker.html]

Frew, James, Michael Freeston, Linda Hill, Greg Jane, Mary Larsgaard, Qi Zheng. (1999). "Generic Query Metadata for Geospatial Digital Libraries." In Proceedings of the Third IEEE Meta-data Conference, April 6-7, 1999 [http://computer.org/proceedings/meta/1999/papers/55/jfrew.htm]

Lagoze, Carl. (2000). *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience* http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1801

Lessig, Lawrence. (1999). *Code and other Laws of Cyberspace*. New York: Basic Books.

Licklider, J.C.R. (1965). *Libraries of the Future*. Cambridge: MIT Press.

EU-NSF Working Group on Metadata. (1999?). *Metadata for Digital Libraries: a Research Agenda*. [http://wwwlis.iei.pi.cnr.it/DELOS/REPORTS/metadata.html]

International Organization for Standardization (ISO). (2000) *Overview of the MPEG-7 Standard* (version 3.0, May/June 2000) ISO/IEC JTC1/SC29/WG11 N3445 [http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm]

Svenonius, Elaine. (2000). *The Intellectual Foundation of Information Organization*. Cambridge: MIT Press.

Notes:

1. Licklider (1965, e.g., p. 39)
2. Although the prediction was accurate in its timing, Licklider envisioned systems that drew much more extensively on the concepts of artificial intelligence being explored in the 1960s than has been the case. Today's information system components for search and retrieval, such as Internet search engines and tools for matching gene sequences and documents, rely heavily on brute force methods made possible by the development of ever faster processors and networks, and ever denser media for computer memory and data storage.
3. Some URLs to try: http://www.badgers.org.uk/; http://www.uclan.ac.uk/library/musrail.htm;

http://skipjack.net/le_shore/worcestr/birding/birding.html

4. In keeping with the usage adopted by the World Wide Web Consortium and the Dublin Core Metadata Initiative, I use metadata as a singular collective noun and the anglicized plural for schema.
5. Svenonius (2000, chapter 1, p. 2)
6. Lessig (1999, chapter 7, p. 87). One of Lessig's main points is that regulation of the architecture of cyberspace is as necessary to society as the regulation of physical space (through building codes, establishment of parks, environmental controls, etc.). The consitution and most existing laws, however, were framed in a world constrained by physical space and demonstrate "latent ambiguities" when applied to cyberspace.
7. EU-NSF Working Group on Metadata. (1999?)
8. Svenonius (2000, chapter 2, p. 17)
9. Svenonius (2000, chapter 2, p. 20)
10. Dublin Core Metadata Initiative. http://www.purl.org/dc/
11. Open eBook Forum. http://www.openebook.org/
12. JPEG2000. http://www.jpeg.org/JPEG2000.htm
13. MPEG-7. http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm
14. IMS Meta-data Specification. http://www.imsproject.org/metadata/
15. Federal Geographic Data Committee. http://www.fgdc.gov/
16. VRA Core Categories, Version 3.0. http://www.gsd.harvard.edu/~staffaw3/vra/vracore3.htm
17. Art Museum Image Consortium (AMICO). http://www.amico.org/
18. Visual Information Access (VIA), Harvard University. http://hul.harvard.edu/ldi/html/via.html
19. Encoded Archival Description.http://lcweb.loc.gov/ead/
20. California Digital Heritage Image Finding Aids, Online Archive of California, California Digital Library. http://www.oac.cdlib.org/dynaweb/ead/calher
21. Rare Book, Manuscript, and Special Collections Library, Duke University. http://scriptorium.lib.duke.edu/
22. International Committee for Documentation of the International Council of Museums (ICOM-CIDOC). http://www.cidoc.icom.org/
23. The Perseus Project. http://www.perseus.tufts.edu/
24. Licklider (1965, p. 5)
25. Draft Standard Z39.85-200X, The Dublin Core Metadata Element Set. http://www.pe[illegible]us.tufts.edu/
26. Lagoze (2000)
27. Dublin Core Qualifiers. http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm
28. Baker (2000)
29. Svenonius (2000, Chapter 7, endnote 12, p. 214)
30. Svenonius (2000, Chapter 6, p. 93)
31. Dublin Core Data Dictionaries, University of Washington Libraries. http://www.lib.washington.edu/msd/mig/datadicts/
32. Frew (1999)

33. Federal Geographic Data Committee. http://www.fgdc.gov/
34. Resource Description Frameowrk. http://www.w3c.org/RDF/

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Thomas A. Downing

Chief, Cataloging Branch
Library Programs Service
United States Government Printing Office
Washington, D.C. 20401

An Initial Survey and Description of How Selected United States Government Libraries, Information Centers, and Information Services Provide Public Access to Information Via the Internet

About the presenter:

Thomas A. Downing has been the Chief of GPO's Cataloging Branch, Library Programs Service since 1992. Prior to this time he held management positions in GPO's Documents Sales Service. He holds a BA in Political Science from Western Michigan University, a Master of Arts in Hebrew Literature and Cognate Studies from Hebrew Union College, and a Master of Science in Library and Information Science from Simmons College. Tad, as he is most widely known, has published articles in such journals as the Journal of Government Information, The Serials Librarian, CONSERLINE, and the OCLC Newsletter. He has represented the National Cataloging and Indexing Program for U.S. Government Publications before national and state library associations and is GPO's representative to CONSER and BIBCO. Tad leads the Cataloging Branch's participation in OCLC's CORC Project and is on the editorial board of The Serials Librarian. His operational interests include identifying and evaluating the most feasible options for providing efficient and effective cataloging services and access to online publications within the context of evolving national standards.

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

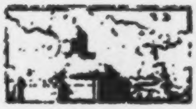
Full text of paper is available

Summary:

The purpose of this survey is to describe how selected United States Government agencies provide information to the public via Internet services. With more than 2,000 Federal library and information centers located throughout the world this effort, of necessity, is selective and findings neither represent all libraries nor do they identify all approaches currently used to present information via the Web.

An effort has been made to describe services without attributing values to particular site characteristics, e.g., bibliographic record applications are not considered superior to browse applications. Those who wish to consider evaluative criteria applicable to such an effort may consult a recently published study entitled Performance Measures for Federal Agency Websites: Final Report, by Charles R. McClure, et.al.

This report provides a brief snapshot in time of a complex and rapidly evolving world. While not definitive in scope, it is hoped that this report will provide a baseline for anyone who may wish to revisit some of these sites in the future to determine how services may have been expanded, reduced, or refined.



Library of Congress
December 18, 2000
Comments: lcweb@loc.gov

An Initial Survey and Description of How Selected United States Government Libraries, Information Centers, and Information Services Provide Public Access to Information Via the Internet

Thomas A. Downing
Chief, Cataloging Branch
United States Government Printing Office

Final version

Purpose

The purpose of this survey is to describe how selected United States Government agencies provide information to the public via Internet services. With more than 2,000 Federal library and information centers located throughout the world this effort, of necessity, is selective and findings neither represent all libraries nor do they identify all approaches currently used to present information via the Web. [i]

An effort has been made to describe services without attributing values to particular site characteristics, e.g., bibliographic record applications are not considered superior to browse applications. Those who wish to consider evaluative criteria applicable to such an effort may consult a recently published study entitled Performance Measures for Federal Agency Websites: Final Report, by Charles R. McClure, et.al. [ii]

This report provides a brief snapshot in time of a complex and rapidly evolving world. While not definitive in scope, it is hoped that this report will provide a baseline for anyone who may wish to revisit some of these sites in the future to determine how services may have been expanded, reduced, or refined.

Methodology

Nineteen sites have been selected for this survey. Sites were selected after being identified through linkage applications for online government information resources or by consulting the United States Government Manual [iii]. I express, in advance, my regret for the unintended omission of sites that would have significantly contributed to the collection of data or to an understanding of how selected Federal agencies provide information to the public via the Internet. Those who wish to provide referrals to additional sites, to correct site information that was gathered before October 27, 2000, or to advise me of changes that have occurred since the 27th may do so by contacting me at: *tdowning@gpo.gov*.

Sites with significant access or security restrictions for most of their applications have not been included in the survey. Several non-intelligence community sites with restricted applications have been included because, in my judgement, substantive applications were accessible.

As might be expected the nineteen sites remaining within this survey are similar in many respects, dissimilar in some respects and many possess notable attributes. Notable or unique attributes not anticipated within the survey checklist are not represented in the matrix but are selectively noted as part of a narrative.

Relying on my eyes and judgement (sometimes enfeebled by staring too long at a screen) I have attempted to provide a singular review of resources that corresponds with a consistent effort to match characteristics with matrix topics. This approach reduces the likelihood that more than one person will see either resources or matrix topics differently and also places responsibility for omissions with one person.

An average of slightly more than one hour has been spent per site. As befits averages, approximately one hour was inadequate for some sites and more than adequate for others. Time spent viewing those sites chosen for the matrix varied from approximately forty minutes to approximately one-hundred-thirty minutes. Times spent viewing sites varied depending upon how many applications were available to review, how easily they were to identify and test and, I admit, on how entertaining (or not) the site was to visit. I believe that I have made reasonable efforts to work through applications to discover subsidiary resources that are not readily apparent from the main page of a site. However, it is possible that some notable applications were not identified because, at the time of viewing, they were deeply imbedded in pages far from the main site page.

Three important factors affecting the quality of site reviews and the use of sites are how readily identifiable, organized, and intuitive site applications are for people to use. No effort has been made to apply these subjective factors for an evaluation of sites. Those who wish to form their own conclusions relative to this survey may access what I consider the main page of each site via hyper links from each institution within the matrix. I thank Mr. Theodore Defosse, of my staff, for his assistance with making this matrix presentable and for creating hyperlinks to site applications.

Background and Rationale to Selected Questions:

A number of questions made part of this survey require explanation. The first two survey questions are intended to determine how many sites describe themselves as "digital" or "virtual" libraries, etc. These are matters of nomenclature and self-definition with no clear meaning as to what a user should expect of sites identified by such terms. Only two of nineteen sites within the survey use either term.

The question, "No collection of online works: search engine only" is intended to identify those sites that provide a search engine only from which to launch Internet inquiries. A search engine only application is seen as distinct from those sites that provide access to a collection of resources via bibliographic records, browse applications, or a keyword search window. In my judgement, only two sites of nineteen consist of a search engine only.

The question associated with archiving of resources for "permanent" access is a very difficult question to answer. With the exception of the Library of Congress and the U.S. Government Printing Office (each with statements indicating archival activities) it is not presently known to what extent other institutions are making efforts to assure permanent public access to online resources. No readily identified statements concerning archives or permanent access to online resources were discovered at other sites.

The portion concerned with "Bibliographic Record Applications" relates to applications that use distinct records to describe resources in bibliographic terms, i.e., title, series, classification, subjects, notes, etc. No effort has been made to determine cataloging standards used in creating records. Similarly, no effort has been made to account for the many differing methods of displaying such information.

The National Transportation Library (NTL) contains a search engine that produces records of resources that correspond more closely with what could be identified as "citations" (basically a search engine generated title) than bibliographic records. In effect, these citations are composed of elements of search results and do not presently represent distinctive bibliographic records. This situation was not anticipated by survey questions.

The question, "Online resource records contain online addresses?" is for those who wish to identify the URL associated with a link in a record so that, if the link is broken, information concerning the most recent link may be available for re-establishing a connection. Some bibliographic applications take users to online resources without an address in the record.

As with the first two survey questions, the question, "Are some resources identifiable via 'Subject Bibliographies'?" is more a matter of nomenclature than of substance. In effect, no distinction is made, except in name, between those browse applications that identify resources by topic and those that do so by the term "subject bibliographies". A "yes" to the former question need not require a "yes" to the later question.

Several questions concern GILS applications. The term GILS has evolved in recent years from representing "Government Information Locator Service" to "Global Information Locator Service". Over time, the use of GILS by Federal agencies to identify and locate their information resources has seemed sporadic and varied. Survey results indicate that, at present, few Federal agencies provide this service in conjunction with web pages associated with libraries, information centers, and information services.

One of the most persistently difficult site characteristics to identify concerns the matter of access to online resources. Many sites provide information about non-Federal resources but, given licensing restrictions, do not provide access to the resources themselves. Given the scope of this survey, which is to determine how sites provide the public with access to information generated by all entities, efforts have been made to determine if sites provide access to more than only U.S. Government resources. Given the considerable time required to thoroughly test access to a wide assortment of resources and the time restrictions associated with this survey, corrections to this segment of the survey are expected.

Questions regarding identification and access to images concern discretely identified collections of images (manuscripts, maps, photographs, etc.). At present, although several sites possess notable image collections, most sites in the survey do not.

Efforts have been made to identify and distinguish those sites that refer users to partner sites (with some formal notice of partnership) from those that refer users to sites without such statements. Referrals for off site resources by Federal agencies, whether to partner sites or not, imply some responsibility to monitor off-site content. Although considered literature, access to Burton's translation of the *Kama Sutra* from a tertiary site pointed to by a surveyed U.S. Government agency may seem inappropriate to some people.

The Survey Matrix

Selected Information Extracted from the Survey:

All nineteen sites contained various statements concerning the scope of the "collection" or of resources associated with services.

All but one site contained information concerning the scope of services provided by personnel associated with the site.

All but one site provided some level of access to online information.

Only two of nineteen sites included statements associated with online archives or services associated with "permanent public access" to online resources.

Data collected indicate the following with regard to types of resources made available:

84.21% of sites (16) provide access to online serials

78.94% of sites (15) provide access to online monographs
78.94% of sites (15) provide access to online databases
31.57% of sites (6) provide access to online maps

Data collected indicate the following with regard to how sites identify available information:

68.42% of sites (13) identify resources via browse applications
57.89% of sites (11) contain a general search window
52.63% of sites (10) identify resources via bibliographic records
52.63% of browse applications (10) are browse titles
47.63% of browse applications (9) are browse topics
10.52% of sites (2) identify resources via GILS
5.26% of browse applications (1) are browse report numbers

Data collected indicate the following with regard to the accessibility of information identified by search applications:

63.15% of sites (12) restrict access to some online resources, including some resources published by some United States Government agencies.

Data collected indicate the following with regard to the presence of "Kid's Pages":

47.36% of sites (9) contain "Kid's Pages"

Data collected indicate the following with regard to online information referrals (links) to other sites:

68.42% of sites (13) link to non-partner U.S. agency sites
63.15% of sites (12) link to non-partner education sites
52.63% of sites (10) link to non-partner commercial sites
42.10% of sites (8) link to non-partner state and/or local government sites
31.57% of sites (6) link to U.S. agency sites identified as partners of the institution

The following selected sites contain what I consider to be notable or unique applications:

Environmental Protection Agency:
Your Community application, which allows users to input a zip code to retrieve information concerning local sites associated with reported violations of environmental laws, rules, and regulations.

GPO Access:
Notable for more than 11,000 bibliographic records with hyper links to online United States Government

resources and for Ben's Guide, with applications that provide information relating to the processes of local, state, and national governments for children of all ages and adults.

Library of Congress:
Hundreds of thousands of images and many sound recordings associated with the American Memory Project and online galleries of exhibitions.

NIST Virtual Library:
Organization charts with links to related personnel, programs and services associated with NIST organizations.

NOAA Library:
Astonishing assortment of images and links associated with climate and weather.

Smithsonian Institution Libraries:
Hundreds of thousands of images of art and artifacts selected from the collections.

Labor Dept. Library andWirtz Labor Library:
Images of labor movement related posters.

Institute of Peace Library:
Links to foreign ministries, governments, and peace related institutes and research centers.

PRELIMINARY CONCLUSIONS:

An initial review of survey data suggests the following:

Most sites use a combination of bibliographic records, browse applications, and search windows for identifying and providing access to online information.

With the exception of two sites, GILS applications are not used for identification/access to information.

No one method, either bibliographic records, browse applications, search windows, or GILS applications account for 100% of the means for identifying or accessing online information. Sites provide users with options for accessing selected types of information and may provide multiple applications based, in part, on differing degrees of labor that are required to create and maintain applications.

At most sites, a user's ability to identify online resources does not guarantee access to resources that have been identified. Although restrictions to access are most commonly associated with commercial information, some services also restrict access to some U.S. Government information resources.

Thoughts Concerning the Future of Online Information Services

Predictions concerning the future depend upon many factors. These include how effectively and persistently the public expresses interests in improving online services, the level of appropriated funds available to U.S. Government agencies for development of online services, and the priority for spending appropriations on online resource applications within individual agencies.

Appropriations aside, it is likely that much of the experimentation that characterized many of the "here today, gone tomorrow" applications at Federal websites during the past six or seven years will be replaced by more stable applications. These methods will reflect an emerging sense of "best practices" for making online information accessible. Applications will continue to evolve and no single method for identifying and accessing information is likely to replace alternative methods. I believe, however, that it is reasonable to make the following predictions:

Regardless of press releases and agency hype, no single information service will be adequate to the task of providing comprehensive, predictable, and authoritative description and access to all online resources published by United States Government agencies. Such a service is possible only with the infusion of massive appropriations to fund an infrastructure that is capable of identifying, describing, and providing access to all known U.S. Government resources.

Efforts to create "one-stop-shopping" sites such as FirstGov.gov, with links to many other sites, may evolve to provide the public with a useful adjunct to locating well established services maintained by major agency providers.

Agencies will continue to improve main pages of websites to provide users with a more intuitive sense of how to identify and access online resources that fall within the scope of agency interests.

An increased number of agencies will apply resources to maintain archival services, to support data migration, and to provide permanent public access to many online works.

The imbedding of metadata (information about information) into online publications for identifying such information as titles, series, etc. will support improved data collection needed for more automated cataloging and locator service applications.

No matter how inaccurate these predictions, readers may be assured that the continued use of the Internet to provide online information will make for few dull moments, both for those who provide online services and for those who use them.

---

[i] *FLICC Factsheet -- Federal Library and Information Centers: A National Resource*. The Federal Library and Information Center Committee, Library of Congress, 2000.

[ii] *Performance Measures for Federal Agency Websites: Final Report to Sponsoring Agencies: Defense Technical Information Center Energy Information Administration Government Printing Office*, by Charles R. McClure, et. al., October 1, 2000.

[iii] *The United States Government Manual 2000/2001*, Office of the Federal Registrar, National Archives and Records Administration, revised June 1, 2000.

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Jane Greenberg

Assistant Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
CB #3360. 207 Manning Hall
Chapel Hill, NC 27599-3360

A Comparison of Web Resource Access Experiments: Planning for the New Millennium

About the presenter: Jane Greenberg teaches in the School of Information and Library Science, University of North Carolina at Chapel Hill (UNC). Her research interests focus on lexical-semantic relationship systems and on metadata and classification problems that involve the organization and retrieval of information objects, including images, archives, and multimedia resources. Greenberg's current metadata research involves serving as a project principal and the Metadata Coordinator for the North Carolina Plant Information Center, a partnership funded by the Institute of Museum and Library Services. Greenberg is a member of UNC's Open Source Research team, where she has examined the production of Linux Software Maps and participated in metadata-based analyses that examine the evolution of the open source community. Greenberg has taught metadata workshops nationwide for AMIGOS, PALINET, SOLINET, and SOASIS. She is on the editorial board of the *Journal of Internet Cataloging*. Prior to earning her doctorate, she held a number of posts as professional librarian and archivist, the most recent of which was as the Coordinator of Special Collections Cataloging at the Schomburg Center for Research in Black Culture, a Research Division of the New York Public Library. Full text of paper is available

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

Over the last few years the bibliographic control community has initiated a series of experiments that aim to improve access to the growing number of valuable information resources that are increasingly being placed on World Wide Web (here after referred to as Web resources). Much has been written about these experiments, mainly describing their implementation and features, and there has been some evaluative reporting, but there has been little comparison among these initiatives. The research reported on in this paper addresses this limitation by comparing five leading experiments in this area. The objective was to identify ***characteristics of success*** and ***considerations for improvement*** in experiments providing access to Web resources via bibliographic control methods. The experiments examined include: OCLC's CORC project; UKOLN's BIBLINK, ROADS, and DESIRE projects; and the NORDIC project. The research used a multi-case study methodology and a framework comprised of five evaluation criteria that included the experiment's *organizational structure, reception, duration, application of computing technology, and use of human resources*. This paper defines the Web resource access experimentation environment, reviews the study's research methodology, and highlights key findings. The paper concludes by initiating a strategic plan and by inviting conference participants to contribute their ideas and expertise to an effort will improve experimental initiatives that ultimately aim to improve access to Web resources in the new Millennium.

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

A Comparison of Web Resource Access Experiments: Planning for the New Millennium

Jane Greenberg
Assistant Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
CB #3360. 207 Manning Hall
Chapel Hill, NC 27599-3360

Final version

Introduction

The exponential growth of the Internet, particularly the Web, has created many new challenges for librarians in their role as collectors, organizers, and access providers of information resources. This is particularly evident in the area of bibliographic control, where librarians are conducting a wide variety of experiments in order to provide access to new information formats and resources that are often volatile. While these experiments have been operational for a few years and they have been evaluated at least on an informal level (see project homepages, URLs given below in the *Method* section of this paper), there is little evidence of their having been compared in any unified way. One reason for this limitation is that these experiments are fairly new and developers have been more concerned with implementation than formal comparisons. It's likely that the innate differences among these experiments has also hampered comparison activities. That is, while these experiments all aim to improve Web resource access by bibliographic means, they differ greatly in their design and features, which makes comparison a difficult task. Despite such observations, these experiments can be compared at least on a general level, and they need to be compared if efforts in this area are to improve in the new Millennium.

Another way to consider this issue is that if researchers and leaders in the bibliographic control community had sufficient knowledge of what characteristics contributed to successful experimentation, they could be continued and incorporated into future projects. Likewise, an agreed upon list-perhaps even an official list-of considerations for improvement could direct future research agendas, encourage the development of alternative and innovative techniques, and ultimately have a positive impact on the next generation of experiments that aim to provide access to Web resources. The multi-case study reported on in this paper addresses these issues by examining five leading Web-based bibliographic access experiments and comparing them in a unified manner.

Experimentation and Web Resource Access

Experimentation in the scientific world involves the use of methodical investigative techniques to examine a particular *problem* or a *series of problems*. The bibliographic control community has responded to the *problem* of resource access that stems from the Web's exponential growth with experiments, such as OCLC's CORC project; UKOLN's BIBLINK, ROADS, and DESIRE projects; the NORDIC project; and a series of other initiatives. These experiments aim to improve the *finding, gathering*, and *evaluating* functions outlined in Cutter's Rules for a Dictionary Catalog (1901), and provide a stimulus for bibliographic control activities in the context of the Web. Several other factors allow these projects to be defined as experiments:

- They have been *implemented under fairly controlled conditions*. The development and implementation each project is managed by a group of persons, such as an advisory board, and project membership is generally geared to a select community or environment.
- They *test new information technologies*. All of the examined projects work with Web-based hypertext, networked communication, and there is evidence of experimentation with automatic computer processing techniques.
- They are *among the first examples their kind*. All of the examined projects test novel procedures and processes. (For example, the Nordic Project is among one of the first large-scale projects to test the implementation of a resource creator metadata template based on the Dublin Core and BIBLNK is among one of the first projects to explore how national bibliographic agencies and publishers can work together to establish authoritative bibliographic information for electronic resources.)
- They *include an evaluative component*. It is impossible to design and implement an experiment without some aspect of evaluation. Each of these experiments has involved at least some form of evaluation--such as testing a design feature or observing usage statistics.

Project-specific evaluation is an important activity and can be critical to the success of these experiments. What is equally important at this stage of Web-based bibliographic control experimentation is research like that presented in this paper, which compares these initiatives.

Objectives of the Study

The study reported on in this paper examined experiments that aim to improve access to Web resources via bibliographic control methods. The objective was to compare these experiments and to identify ***characteristics of success*** and ***considerations for improvement***. The following two research questions guided the study:

1. What characteristics are found in successful Web resource access experiments?
2. How can these Web resource access experiments be improved in the future?

Method

The investigation was a muti-case study that compared five experiments developed to improve access to Web resources via bibliographic control processes. The multi-case study method was selected because it was the best way to observe relationships among these experiments. Experiments examined include:

1. BIBLINK: Linking Publishers and National Bibliographic Services [http://hosted.ukoln.ac.uk/biblink/]
2. DESIRE (Development of a European Service for Information on Research and Education) [http://www.desire.org/]
3. Nordic Metadata [http://www.ilrt.bris.ac.uk/roads/]
4. OCLC CORC (Cooperative Online Resource Catalog) [http://www.oclc.org/oclc/corc/]
5. ROADS (Resource Organization and Discovery in Subject-based services) [http://www.ilrt.bris.ac.uk/roads/]

A table outlining project goals and status is found in Appendix A.

Evaluation Criteria

A framework comprised of five evaluation criteria served as a basis for the study and allowed the experiments to be compared in a unified manner. These are defined as follows:

1. Organizational structure.
   The experiment's structural foundation defined by its goals, administration (project leaders, members, and partners), and funding.

2. Reception.
   The experiment's acceptance by the professional information community (e.g., librarians and other information professionals) and the larger general public.

3. Duration.
   The experiment's time expanse and indicators of progression (e.g., alpha and beta release, version number, or phase).

4. Application of computing technology.
   The experiment's exploitation of computing technology.

5. Use of human resources. The experiment's ability to harness and optimize human knowledge and

- skills.

The criteria framework allowed for the five experiments to be studied in a unified manner despite their differences. The procedures were criteria centric, in that each criterion was examined *one-at-a-time* across all five experiments before the next criterion was studied.

Results

The multi-case study permitted the identification of *characteristics of success* and *considerations for improvement* in experiments that use bibliographic control methods to improve access to Web resources. These results are presented below within the context of the evaluation criteria underlying the study.

Characteristics of Success

Determining the overall success for each experiment requires an in-depth analysis beyond the scope of a single paper. This study took a more general approach and compared these experiments at a higher level in order to identify characteristics of success that were applicable to all of the projects. These characteristics can be discussed under the framework of the five evaluation criteria used in this study.

Organizational structure

The study revealed a number of similarities across the experiments that appear to have contributed to their success. To begin with, each experiment is defined by a list of *goals*.[1] Clearly, goals alone do not guarantee success, and it is recognized that goals may change throughout the experimental process. What is important here is that the goals provide a focused direction and appear to contribute to a successful experiment. Related to goals, each experiment has a defined administrative structure comprised of project leaders, members, and/or partners-and to take this observation a step further, project participants were found well beyond the confines of a single institution. It seems that an obvious *administrative structure*, particularly one overseeing decision-making processes, and that partnerships beyond the confines of a single institution may both be characteristics of successful experimentation. A final factor under this criterion that, no doubt, contributes to successful experimentation is adequate *funding*. Government funding supported four of the five experiments, and CORC is funded by its members, which mainly includes libraries.

Reception

How well a development is received by a community can be an indicator of success. The mutli-case study was enhanced with an electronic survey that was sent to five bibliographic control professionals, five information professionals who are not engaged in bibliographic control, and five general Web users.[2] The survey asked participants if they had knowledge about any of the five experiments examined in this study, if they had knowledge of Yahoo! and Lycos, and what was their preferred starting point for a Web search. While the participant sample was convenient, and not necessarily statistically sound, the results in conjunction with the results of the multi-case study are helpful in examining this criterion.

A majority of the bibliographic control professionals were aware of more than one of the experiments evaluated in the multi-case study, and three of the five persons in this group referred to the availability of what they called *excellent* documentation and tools. For example, the NORDIC Dublin Core metadata templates (http://www.lub.lu.se/cgi-bin/) and the corresponding User Guidelines for Dublin Core Creation (e.g., http://www.sics.se/~preben/DC/DC_guide.html). All five of the experiments evaluated have fairly substantial documentation-and in a number of cases support access to tools. It is likely that open documentation and access to tools contribute to the success of an experiment.

Duration

The average duration of the experiments examined for this study is three years. This is a result of the fact that the three UKOLN experiments (BIBLINK, DESIRE, and ROADS) had defined time frames in which they successfully completed designated tasks, and that the Web is less than a decade old. The NORDIC project is in its second phase, which began in January 1999, and that CORC, which was launched in Janurary1999, is a fully functional project that is no longer considered experimental. The progression from alpha and beta testing, and/or through various phases or versions is demonstrative of success. What is particularly exciting in the realm of duration is the model offered by CORC, which promotes continued growth of the experiment via its transition from the experimental stage to a fully operational project. Related to this observation, however, one must consider the result of the experiments that may have had a shorted time frame, but have long-term impact. For example, the ROADS project software toolkit is still accessible and continues to be used in various services like the Social Science Information Gateway (SOSIG) and the OMNI Health and Medicine Gateway (e-mail correspondence with Michael Day, Research Officer, UKOLN The UK Office for Library Information, University of Bath).

Application of computing technology

Today, it is impossible to discuss bibliographic control and computing technology without referring to online catalogs and the MARC format. These developments take advantage of computing capabilities to support *interoperability amongst information systems, expedient and efficient resource organization and access*, and *distributed cataloging* via networked communication protocols. The experiments examined for this study have successfully taken advantage of Web-based technology in a similar way in order to support *resource discovery and communication* among different institutions. Beyond these developments, computing technology offers many more sophisticated capabilities, particularly in the area of information retrieval. The ROADS project has been among one of the most successful projects in this area, promoting searching across multiple gateways, harvesting, relevance ranking of retrieval results, interface customization, hierarchical browsing, and multi-lingual access (ROADS User Survey Results, 1999). Another example is found with the CORC project, which includes the Scorpion algorithm (Shafer, 1998) for automatic classification.

Use of human resources

The final evaluation criterion involves the use human resources. Despite great strides by the artificial intelligence community, human beings can still outperform computers in *most* complex intellectual tasks. This coupled by the fact that networked protocols can facilitate communication and can unit the talent and skill of persons involved in the production, acquisition, organization, and life of a Web resource invites new documentation possibilities. The experiments reviewed in the multi-case study involve administrators, bibliographic control and other information professionals, and in some cases Web resource creators (document authors). The NORDIC project and CORC both allow for metadata to be created by professionals and resource creators, as long as there is quality control by professionally trained metadata experts. BIBLINK involves another collaborative relationship in that national bibliographic agencies work with publishers to establish authoritative bibliographic information for electronic resources. These partnerships are successful in that they harness and optimize the expert knowledge of bibliographic control professionals by having them focus more-or-less exclusively on activities that require their expertise, while persons or agencies with less skill are responsible for the simpler yet time consuming bibliographic control tasks.

Considerations for Improvement

While the evaluation of Web resource access experiments allowed for the identification of characteristics of success, it also permitted the identification of project features and aspects that could be improved in future initiatives. These considerations for improvement can also be discussed under the rubric of the five evaluation criteria underlying the study.

Organizational structure

The organizational structure as noted above seemed to be sufficient for the current undertaking of experiments. What needs to be considered now, however, is long-term experimentation as viewed with initiatives such as CORC and its progression from an experimental state to a fully operational and growing project. Also important to consider is the funding and membership structure of CORC, which is slightly different than the other experiments. CORC is supported by a source of consistent funding from its members and it involves more partners compared to any of the other experiments examined. A consideration for improvement is to design experiments that include a plan for continued funding and partnerships that extend beyond a single institution.

Related to these considerations is a perceived need for these experiments *to talk to each other* and interoperate, particularly in cases where they are using the same features, or where one feature could be enhanced by another feature. There is evidence of interoperability among several of the partnerships supported by the UKOLN projects, but there is a great deal of room for growth in this area. For example, BIBLINK, CORC, and the NORDIC project all work with a variants of the Dublin Core, but they each have their own implementation. ROADS has developed sophisticated subject gateway that could potentially be used to access CORC and other initiatives. The various experiments in this area could have an even greater impact on Web resource access if they communicated not only on a goal level, but also on an operational level. In other words, a framework is needed that will improve interoperability and permit these experiments to talk to each other more than is currently practiced. Along these lines, experimental initiatives might even consider talking to commercial enterprises, such as search engines, that facilitate access to Web resources-a point that is considered further under the next criterion.

Reception

With respect to the study's supplemental survey on Web searching, the bibliographic control professionals were aware, at least by name, of the majority of the Web resource access experiments evaluated in this multi-case study. However, this group of persons represent a minutely small segment of the Web user population. Of the five information professionals not engaged in bibliographic control activities, only two reference librarians were aware of CORC, the rest of the information professionals and all of the average Web users had not heard of any of the other experiments evaluated. These results were offset by the fact that all of the participants had knowledge of Yahoo! and Lycos, and they named various commercial search engines as their primary means of Web access. The point to consider here is that the larger universe of Web users may use commercial search engines for searching not only because of convenience, but because and they are unaware of Web resource access experiments. This should be of concern to the bibliographic community because these experiments can be costly and labor intensive, and more importantly because it is likely that these experiments will yield far superior retrieval results in various domains. (A comparison between the Web-based bibliographic control experiments and commercial search engine algorithms is beyond the scope of this paper. Additionally, this author believes that access via both mechanisms should not be looked at as being diametrically opposed.)

This said, the limited knowledge about the examined experiments may in part be due to the their experimental nature and short lived status. Also, it is likely that place of origin has had an impact on the general knowledge about these experiments, as only one of the five projects was initiated in the United States. Even so, it is important to remember that Yahoo! Lycos, MOSAIC, and many other Web developments in the United States began as experiments at institutions of higher learning, and are now extremely popular on an international scale.

The bibliographic control community needs to consider building stronger public relations and advertising to populations beyond the bibliographic control community, particularly if these experiments are to thrive. *Documentation and tools* supporting the experiments examined in this study were attractive to bibliographic control professionals, but it seems these resources may not really foster or invite project exploration or use for information professionals not engaged in bibliographic control as well as for the general Web user. Perhaps the bibliographic control community should explore advertising practices, or some variant of this activity, as viewed in the commercial sector. Along these lines, Web resource access experiments might even consider collaboration with commercial search engines or other for-profit initiatives in some form. A partnership with a commercial enterprise, no doubt, requires serious exploration, but it is not unrealistic to ask conference participants to think about this question, especially when it is known that a segment the Northern Light search engine/index is using a version of the Dublin Core and commercial search engines are increasingly interested in the application of bibliographic control methods and classification activities, which require the talent and skill of persons trained in bibliographic control methods.

Duration

It is human nature to equate longevity with success. The printed monograph is one the most successful technological innovations, which, despite all forecasts of its demise in the electronic era, is increasing in number yearly (About Book Title Production, 2000; Library and Information Statistics Tables, 1998). As already indicated, three of the five experiments examined have been completed. Their intent was to investigate various aspects of Web resource access in a certain time frame. But their short-lived lives may be used to raise questions about their long-term value. Again long-term value of these project must be considered and the results leading to new research. A case in point is the RENARDUS project (http://www.renardus.org), an outgrowth of the DESIRE project, which is establishing an academic subject gateway service with integrated access and plans to be a long-term venture. In sum, a consideration for improvement under this criterion is to develop and implement experiments that have long-term goals, and which aim to become fully operational projects that support Web resource access.

Application of computing technology

Computing technology surpasses the human in terms of speed and consistency and supports a wide variety of automatic techniques that can be incorporated into and strengthen bibliographic control operations. Examples of these tasks include *natural language processing, automatic classification and indexing, automatic metadata generation* and *searcher profiling*. While the experiments examined in this study explore the use of computing technology, most notably the ROADS project, they do not fully incorporation or exploit automatic processing capabilities. Bibliographic control initiatives need to further explore how to take advantage of and make use of computing technology for its own sake, and also because only through such efforts can human resources, the last criterion, be fully optimized.

Use of human resources

As indicted in the discussion on characteristics of success, these experiments involve administrators, bibliographic control and other information professionals, and in some cases Web resource creators (document authors), and several initiatives have established collaborative partnerships among persons that have varying skills. There is, however, much room for growth in this area, particularly with communication options that are facilitated by networked protocols. Along these lines, the bibliographic control community needs to identify what tasks can be *accurately, efficiently*, and *superiorly* performed by the computer, but also what tasks ***need*** to be performed by people-and specifically who should perform such tasks so that the bibliographic control professional's expertise can be fully taken advantage of in the aim for access to Web resources.

Call for a Strategic Plan

The bibliographic control community has responded to Web's exponential growth with a series of experiments that aim to improve access to Web resources. These experiments are important because they use traditional cataloging and classification practices and test innovative ideas, processes, and features in environments that extend beyond the library catalog. While these initiatives all aim to improve Web resource access, they differ in some very fundamental ways. Even so, these experiments can be compared, and it is through this type of analysis that the bibliographic control community can identify characteristics of success, considerations for improvement, and initiate superior Web resource access experiments. This paper concludes by suggesting five agenda items that will serve as a base for a strategic plan in this area and by inviting conference participants and other readers of this paper to contribute their ideas and expertise to an effort that will improve experimental initiatives that aim to improve access to Web resources in the new Millennium.

1. Explore considerations for improvement identified via the multi-case study reported on in this paper. These items include exploring the following:

    - secure continued and substantial funding options,
    - explore new and larger-scale collaborations-even partnerships with commercial enterprises,
    - increase public relations and advertisement to communities beyond bibliographic control professionals,
    - plan for long-term experiments and their transition to a fully operational projects (e.g., CORC),
    - facilitate and test interoperability, so that these initiatives can really talk to each other,
    - exploit computing technology and incorporate automated processes into bibliographic operations in an intelligent manner, and
    - optimize the knowledge, skill, and other talents of available human resources.

2. Continue to evaluate projects
The specific features and practices supporting the experiments examined in this study and in other initiatives need to be researched on an in-depth level, and these projects need to be compared to each other on an array of levels. Moreover, efforts should be made to employ scientific research methods to such investigations to insure the constructions of a sound body of knowledge. Only through such research efforts can a pool of knowledge be developed to improve future Web resource access experimentation and access to Web resources.

3. Share research
All of the experiments examined in this study have conducted evaluations, at least on an informal level. The results of these undertakings appear to be accessible via most project Web pages, but they are not generally disseminated to the larger bibliographic control community through professional and scientific publications and conferences. A central vehicle of communication, such as an electronic bulletin board or a Web site is needed so that the results of all research efforts, both formal and informal, can be shared with the bibliographic and related communities. This type of central sharing ground would greatly assist future initiatives, allow for timely access

to results and lessons learned, and permit meta-analyses so that superior Web resource access experiments could be conducted.

4. Develop an official list of considerations for improvements
The bibliographic control community should support the construction of an official list of features, applications, and other aspects that could improve Web resources access experiments that employ bibliographic methods. The research conducted here may serve as a starting point, but an official list would assist the larger bibliographic control and information community, and ultimately help to direct future research agendas.

5. Develop a master Request for Proposal (RFP) for Web resource access experimentation
The last strategic step to be suggested in this initial draft is to develop a master RFP that could direct Web resource access experimentation. The bibliographic control community has developed master RFPs for online catalogs to share, as demonstrated by CONDOC (Crosby, 1997). A master RFP that recommends an organizational structure, outreach and duration plans, and the best way to use of computing technology and human resources could greatly assist institutions and persons at all levels who want to conduct experiments that aim to improve access to Web resources via bibliographic control methods.

Conclusion

This paper defined the Web resource access experimentation environment and reported on the results of a mutli-case study that compared a series of experiments that are innately different, but which all aim to improve access to Web resources. The Web, while increasingly perceived as a trusted vehicle for the dissemination and recording of information, is still very much in a developmental stage. In fact it has been predicted that by the end of the first decade of the new Millennium, the Web will look vastly different and even unrecognizable compared to today. Whether such forecasts will prove true is difficult to gage, but what seems certain is that bibliographic control methods have a role in the new Millennium. The phenomenal growth of the Web has generated a lot of experimentation, but left little time for formal evaluation these initiatives. The bibliographic control community must rise to the challenge and encourage and conduct evaluations so that bibliographic control experiments are successful and assist with the organization and access of information resources that help to of the define the great domain known as the World Wide Web.

References

Crosby, E. (1997). Towards 'CONDOC 2': identifying new requirements for online catalogs. ALCTS Newsletter 8 (3): A-D.

Cutter, C. A. (1904). Rules for a dictionary catalog, 4th ed., (rewritten). Washington, D.C.: Government Printing Office.

ROADS User Survey Results. (1999). Available at: http://www.ilrt.bris.ac.uk/roads/questions.results

About Book Title Production. (2000). Available at: http://publishing.about.com/arts/publishing/gi/dynamic/offsite.htm?site=http%3A%2F%2Fwww.ipa-uie.org%2Fstatistics%2Fannual_book_prod.html

Library and Information Statistics Tables. (1998). Available at: http://www.lboro.ac.uk/departments/dils/lisu/list98/pub.html

Shafer, K. (1998). A brief introduction to Scorpion. Available at: http://orc.rsch.oclc.org:6109/bintro.html

Acknowledgements

The Author would like to thank the following people for comments on this paper, and interest in these experiments: Michael Day, UKOLN Office for Library and Information Networking, University of Bath; Dr. Brian Sturm, School of Information and Library Science, University of North Carolina at Chapel Hill; and Dr. Mary S. Woodley, Ph.D. Social Sciences Librarian California State University. Thank you also to conference organizers and sponsors for fostering such an important dialog.

Notes

1. Note that the experiments are discussed in the present tense in this paper for the purpose clarity. Three of the five experiments have been completed fairly recently, but two are still operational.
2. A convenient sample was used for this part of the study. The sample consisted of five catalogers/metadata professionals, five information professionals (an archivist, a data base administrator, two reference librarians, and a slide curator), and five average Web users (two undergraduate students, an environmental scientist, a professor in education, and an office assistant).

Appendix A

Experiment Goals and Status

Project Name	Project Goal	URL	Beginning Date	Phase or Version No.

DESIRE (Development of a European Service for Information on Research and Education)	"enhancing existing European information networks for research users across europe through research and development in three main areas of activity: Caching, Resource Discovery and Directory Services."	http://www.desire.org/	1996 (Phase I) July 1998 (phase II)	2 phases; completed June 2000
BIBLINK: Linking Publishers and National Bibliographic Services	"to establish a relationship between national bibliographic agencies and publishers of electronic material, in order to establish authoritative bibliographic information that would benefit both sectors."	http://hosted.ukoln.ac.uk/biblink/	Apr. 1, 1996	2 Phases; completed Feb. 15, 2000
ROADS (Resource Organisation and Discovery in Subject-based services)	"1. to produce a software package which can be used to set up subject-specific gateways 2. to investigate methods of cross-searching and interoperability within and between gateways 3. to participate in the development of standards for the	http://www.ilrt.bris.ac.uk/roads/	1995	Completed

	indexing, cataloguing and searching of subject-specific resources"			
Nordic Metadata	"1. enhancement of the existing dublin core specification 2. creation of dublin core to marc converter. 3. dublin core user support and tools evaluation. 4. maintenance and development of metadata tools"	http://linnea.helsinki.fi/meta/	January 1999	II
OCLC CORC (Cooperative Online Resource Catalog)	"to assist libraries in providing their users with well-guided access to web resources"	http://www.oclc.org/oclc/corc/	January 1999	open participation

*Thank you to Paulina Vinyard, Master's Student, School of Information and Library Science, University of North Carolina at Chapel Hill, for her assistance in the compilation of this table.

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Karen Calhoun

Director, Central Technical Services
107-D Olin Library
Cornell University
Ithaca, NY 14853

Redesign of Library Workflows: Experimental Models for Electronic Resource Description

About the presenter:

Karen Calhoun, M.S., M.B.A., is the Director of Central Technical Services at Cornell University Library, a position she has held since March 2000. Prior to that she was the head of cataloging. Active in the development of Cornell's Library Gateway (http://campusgw.library.cornell.edu/) and a frequent speaker on technical services in the digital library, Karen's recent research and operational interests have focused on the organization of networked resources and services, user needs, project management, library workflows, cross-functional teams, and cooperative cataloging and authority control. Currently she leads Cornell's participation in the CORC project, chairs the Program for Cooperative Cataloging (PCC) Standing Committee on Automation and the ALCTS CCS Policy and Research Committee, is active in the PCC Task Group on Journals in Aggregator Databases, and serves as assistant editor of Library Collections, Acquisitions and Technical Services. In addition she has co-edited a special issue of the Journal of Internet Cataloging (forthcoming) on CORC. Before coming to Cornell she held positions at OCLC and the University of Oregon.

Full text of paper is available

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

This paper explores the potential for and progress of a gradual transition from a highly centralized model for cataloging to an iterative, collaborative, and broadly distributed model for electronic resource description. The author's purpose is to alert library managers to some experiments underway and to help them conceptualize new methods for defining, planning, and leading the e-resource description process under moderate to severe time and staffing constraints. To build a coherent library system for discovery and retrieval of networked resources, librarians and technologists are experimenting with team-based efforts and new workflows for metadata creation. In an emerging new service model for e-resource description, metadata can come from selectors, public service librarians, information technology staff, authors, vendors, publishers, and catalogers. Arguing that e-resource description demands a level of cross-functional collaboration and creative problem-solving that is often constrained by libraries' functional organizational structures, the author calls for reuniting functional groups into virtual teams that can integrate the e-resource description process, speed up operations, and provide better service. The paper includes an examination of the traditional division of labor for producing catalogs and bibliographies, a discussion of experiments that deploy a widely distributed e-resource description process (e.g., the use of CORC at Cornell and Brown), and an exploration of the results of a brief study of selected ARL libraries' e-resource discovery systems.

Library of Congress
December 14, 2000
Comments: lcweb@loc.gov

Redesign of Library Workflows: Experimental Models for Electronic Resource Description

Karen Calhoun

Director
Central Technical Services
Cornell University Library

Final version December 2000

Contents:

Looking Back: Technical Services as We've Known Them
Catalogs versus Bibliographies
A New Information Space: the Influence of the Internet and Licensed E-Resources
Space Walks: Heroic Accommodations in ARL Libraries
Towards New Models for Resource Description
Selected Experimental Models
Discussion
Looking Forward: Choosing Which Problems to Solve
Conclusion
Endnotes

Looking Back: Technical Services as We've Known Them

Ten years ago, Younger and Gapen described their vision of technical services in the year 2001. They predicted a paradigm shift characterized by a renewed focus on user needs, client-centered organizational structures, and the merging of technical and public services departments. (1) It hasn't happened. Instead, in academic libraries I know and have known, the most common organizational design remains a functional structure that uses the selection, acquisition, organization, and dissemination of library materials as the basis for logically grouping people and work. The main departments in academic research libraries continue to be collection development, technical services, and public services. However, in the past ten years an upstart department-library systems-has made the traditional library triad into a quadriad (Beile and Adams 2000). (2)

As noted by Stueart and Moran (1993) in their text on library management, a functional organizational structure has distinct advantages: it groups people and tasks that are similar, allows for specialization, and keeps library administrators keenly aware of the contributions and needs of each group. The functional organization's disadvantages include the competition that inevitably arises among departments; a focus on departmental rather than library-wide issues and goals; and difficulty collaborating across departments. (3)

Within technical services, those who organize and process materials for patron use-catalogers-both possess and take great pride in their in-depth knowledge of their specialization-resource description, better known as cataloging. Their chief product is the library catalog, and the process of building and maintaining it has been highly centralized within technical services departments. Library catalogs have served library staff and users well.

The centralized technical services concept has not been without its critics. Younger and Gapen note that in centralized departments "there is too much attention directed toward library processing activities with insufficient focus and attention on meeting users' needs." (4) Noting that "our existing structures are no longer adequate to manage a digital or combination digital/traditional library," Boissonnas (2001) argues for "the deep integration of technical with other reader services" to prepare for the future and to overcome the "fragmentation, overspecialization, and philosophical inertia" associated with the functional division of library work. (5)

Boissonnas, Younger and Gapen's arguments are compelling, and they point to ways in which libraries must change. Nevertheless I maintain that the organizational structures and workflows that have brought us to this point in the development of present-day library catalogs have generally been efficient and effective. Tremendous waves of change have swept over cataloging departments-among them, the advent of shared online cataloging systems, the shift from card to online catalogs, strong growth in special or non-book collections, downsizing, outsourcing, and significant shifts in what professional catalog librarians are expected to do (or not to do). Catalog librarians and their managers-while they have not always done it with tremendous grace-have coped with every wave. That they have managed to do so is an important but largely unacknowledged miracle of the profession. Along the way, they have produced millions of catalog records for their own and the world's libraries, saving millions of dollars by reducing redundant effort by sharing their records in the bibliographic utilities.

The work to reap the full benefits of cooperative cataloging has continued in the past decade, resulting in new initiatives such as the Program for Cooperative Cataloging (PCC). Thomas and Younger (1993) spoke for many when they stated their vision for shared cataloging: "to put into place the necessary support to catalog, once and only once, every item owned or made accessible by libraries and to share that information with all others who need it." (6)

Today, despite the dramatic waves of change I've noted here, and in spite of the exhortations of visionaries, the basic approach to cataloging has not varied. Cataloging departments in academic libraries (and the people in them) remain much the same in terms of organizational structure, role and purpose in the library, understanding of the principles of the catalog, and professional values. Yes, new tasks and responsibilities have been taken on, and continual process improvements have been sought and made, but as overlays or add-ons to the same centralized service model as before. Outsourcing, for example, which provides an external source of catalog records, is by nature compatible with the kind of cataloging that would be done in-house. The PCC, while it represents an essential next step for cooperative cataloging as we have know it, continues to focus on a single set of library-centric resource description standards (AACR2 and MARC) and to not just rely on, but reinforce current library organizational structures. Surprisingly like the academic library reference service departments described by Ferguson (2001), cataloging "remains structured steadfastly around physical objects and the library as place." (7)

Our past experience of technical services in libraries is a powerful lens on how we see the present and the future. These experiences, together with the natural tendency to think the future is going to be like the past, can lead to such strong preconceptions about what resource description is, and who can and should do it, that we ignore critical facts that are in some way external to our mindsets. One such set of critical facts has to do with the history of the division between catalogs and bibliographies.

Catalogs versus Bibliographies

In his 1992 monograph on redesigning library services, Michael Buckland includes a chapter on bibliographies and catalogs.(8) Both bibliographies and catalogs contain resource descriptions, (9) albeit done according to varying standards and conventions, and both are forms of bibliographic control. Arguing that library cataloging can be viewed as a special case of bibliography, Buckland notes that the catalog is like a bibliography in that it is composed of information about works and editions of works. Unlike a bibliography, however, the catalog also concerns itself with individual copies of works-that is, the particular copies that a library holds. He goes on to point out the catalog's usual focus on one particular level of description-for monographs, the edition; and for serials, the title.

By contrast, bibliographies commonly list works at many levels of description (e.g., not only books and serials but also individual journal articles and conference papers). Bibliographies are generally the domain of not only reference librarians, but also individuals and groups operating outside librarianship, such as scholars, professional and scientific societies, government agencies and publishers. In particular, the role of large-scale indexing and abstracting of articles in periodicals has by tradition been left to publishers such as H.W. Wilson.

Buckland hastens to stress that the policy of excluding analytics from library catalogs is a matter of library tradition, not of principle. He concludes that the fact that the catalog is not normally thought of as bibliography "is largely an accident of semantic custom and of a tradition in library organization that associates the catalog with catalogers ... and bibliography with reference librarians."(10)

If one takes a long view of bibliographic control practices and history, then, the responsibility for resource description has been distributed among different groups inside and outside the library for a long time. Yet the boundaries between groups have been drawn so clearly, and the traditional arrangement has worked so well, that many librarians no longer recognize that the present division of labor is only one option among many for getting the work done.

A New Information Space: the Influence of the Internet and Licensed Electronic Resources

Academic libraries have no choice but to respond to the technology and applications of the Internet. Lubans' series of studies of student Internet use at Duke University has documented university students' "growing, even escalating use of Internet resources." (11) Students want the library to offer more Internet-based services. In my own research with Zsuzsa Koltay (Calhoun and Koltay 1999) into users' perceptions of the Cornell Library Gateway, I was struck by a student's comment: "The Gateway is the best [information system] I've ever used, but it is less than optimal. It is a great mock-up of the future." (12) Participants in the Cornell user study wanted the library to continue to add e-resources, especially full text; better communicate with them about the library's networked resources; provide multiple ways of discovering e-resources; and help them help themselves.

In 1992 Buckland proposed a redefinition of the catalog that foreshadowed the explosion of Internet resources. He argued for linking the information in online bibliographies with library holdings and permitting extended searches of multiple bibliographies and catalogs using multiple retrieval systems. (13) By 1999, Van de Sompel and Hochstenbach were experimenting with linking related information entities of all types-citation databases to catalogs and full text, finding aids to primary sources, catalog records to book reviews and images, and more. (14)

Clearly, the future of technical services is a future with Internet technology, applications, and resources in it; but from a technical services perspective, Internet resources tend to break the mold. They cause problems. They challenge our notions of the form and function of the catalog. The rules for cataloging them are in a nascent state. Our present exacting cataloging methods are too slow to handle their volume and complexity in reasonable turnaround times. They change so often that they overwhelm our capacity to maintain them. They force us to question conventional assumptions and workflows.

In keeping with the analysis laid out by Ercegovac (1997) (15), I suggest that Internet resources are driving fundamental changes that demand new operational and organizational assumptions about bibliographic control. The new assumptions are outlined in Table 1.

Table 1. Working Assumptions for Bibliographic Control

NOW	EMERGING
• Local collection, mostly print	• Many kinds of data sets, local and remote
• Catalog represents the collection but is separate from it	• With full text, catalog and collection are converging
• Highly standardized bibliographic records in library schemes (AACR2, MARC format, LC or DDC class, LCSH, Sears)	• Less structure in indexing, mixed representations of data; metadata can be prescribed by varying rules or be free form
• Centralized responsibility for resource description/metadata creation and limited decentralization for specific subjects or languages	• Highly distributed responsibility for resource description/metadata creation; records come from multiple sources

Operationally, electronic resources drive the catalog away from bibliographic control of a physical collection toward the representation and control of a virtual repository and the possibility of a new catalog as described by Buckland. (16) Organizationally, the proliferation of Internet resources is causing a technical services identity crisis.

Most libraries' technical services departments reflect policies and practices that are outgrowths of functional organizational structures. Technical services departments tend to be staffed with individuals who are hired and trained to be experts in some aspect of the acquisition or organization of library materials. As is typical of experts within any type of organization (Neuhauser 1988), it is not unusual for these individuals to be tightly focused on the tasks they perform, to have minimal contacts outside technical services, and to be unfamiliar with the library activities outside their particular function. (17)

This organizational structure has worked well for print resources-selecting, acquiring and describing them is nearly always a linear, sequential process in which one person or group works independently on each step. This is possible because policies and rules are well established and known to participants. But, as can be seen from the workflow description in Table 2, selecting, "acquiring," and describing Internet resources, can be (and often is) an iterative, highly collaborative, looping process that can involve many individuals from many functional groups. The outputs of the process can also vary-the end result may be MARC records in the catalog, links and summary descriptions on a library Web page, or even records in a non-MARC metadata format (Dublin Core or some locally developed record format).

Table 2. "Typical" Progress of a New Electronic Resource

	Steps
1	A selector identifies and selects an electronic resource
2	The selector initiates a request to acquire and/or describe the resource (and/or list it on one or more library Web pages)
3	Acquisitions/selectors/information technology/catalogers/reference staff exchange inquiries as needed
4	Acquisitions or collection development or reference staff negotiate with vendor/publisher/author (for licensed resources)
5	Acquisitions initiates request to describe the resource and/or add it to appropriate Web list
6	Acquisitions/catalogers/selectors/information technology/reference staff exchange inquiries as needed
7	Catalogers consult the resource, resource description standards and databases to prepare resource descriptions (however in some academic libraries it is more common for a resource description to be added to one or more Web lists or databases than for it to be cataloged)
8	Another round of inquiries as needed
9	The resource description is added to the catalog, and/or it gets listed on one or more library Web pages, and/or included in a locally-created searchable database of electronic resources

Space Walks: Heroic Accommodations in ARL Libraries

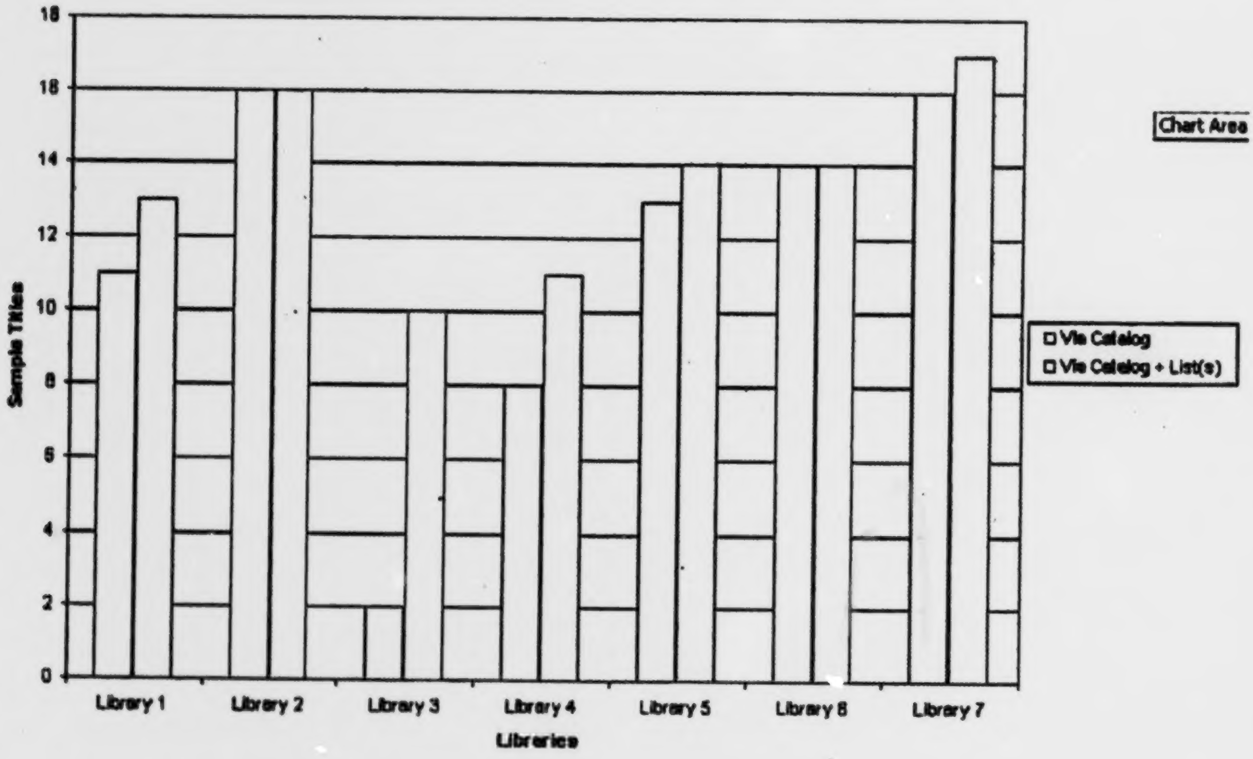
There are countless variations on the "typical" process sketched in Table 2, because libraries have chosen varied models for organizing themselves to manage electronic resources. In an attempt to get a clearer picture of how libraries are accommodating Internet resources now, I completed a brief analysis of the resource descriptions provided for a set of seventeen commonly-licensed online databases and full text journals at seven of the largest ARL libraries in the United States. (18) I did not include titles of e-books in my analysis, although they are approaching us so rapidly now, because how to provide resource descriptions for e-books is still under intense discussion in libraries.

There are two models of networked resource description in operation among the ARL libraries examined. All of the sample libraries provide for discovery and access of online databases and electronic full text journals both via the catalog and via the library Web site. In many cases it appears that collection development or public services is responsible for mounting and maintaining the Web list or lists of titles, while records in the catalog continue to be the domain of technical services.

The format and number of the Web lists vary widely from library to library. One library maintains a searchable database of its networked databases and full text journals that serves as a single point of entry; there are no separate lists, although the contents of the searchable database may be browsed by broad subject category. Two other libraries maintain searchable lists; one is a searchable list of e-journals (no databases) and the other is a searchable list of databases (no e-journals). All of the others maintain lists of networked resources that can be browsed either alphabetically or by broad subject categories. In most cases there are two lists, one for databases and one for e-journals, but in a few cases there are multiple lists of both, and one needs to know which list to pick (e.g., the science and technology one or the social sciences one) before beginning to browse for a particular title.

Figure 1 provides a graphical view of the number of titles that I was able to discover via the catalog alone, or by a combination of using the catalog and available Web lists, for the seven libraries examined. Using the catalog alone, I was able to discover about nine titles, or 54%, on average. (19) The minimum number of titles discovered via the catalog alone was two, or about 12%; the maximum sixteen, or 94%. Adding in what can be found using both the catalog and Web lists, the average number of titles discovered rises to about fourteen (82%), the minimum to eleven (65%), and the maximum to seventeen (100%).

Figure 1. Discovery of Sample Networked Resources in Seven ARLs



The findings provide strong evidence that different functional groups in ARL libraries are already distributing the work of creating resource descriptions for electronic resources. These librarians are making heroic accommodations for Internet resources and adapting their methods to include them. Nevertheless, the traditional boundaries among the functional groups appear to remain intact, although sometimes blurry; few or no libraries appear to be starting over with new service models or examining basic assumptions of specialization. I base this conclusion on my own experiences working with library Web sites and catalogs, conversations with colleagues, and the specific findings of this analysis. Library users probably do want and need multiple ways to discover networked resources. Yet the reality suggested by this analysis is that searchers often must use both the catalog and library Web lists to discover what the library makes available, and even when they do use both, they do not always get the full picture.

This analysis and my own experience further suggest that libraries' current methods for producing electronic resource descriptions generally result in suboptimal, fragmented discovery and retrieval systems that are difficult for library patrons to understand and use. The double work that library staff are often expending to provide multiple access methods-via the catalog and lists-is generally is not paying off as it could, because the efforts by different functional groups tend to be uncoordinated and poorly integrated.

At a minimum, we must redesign and integrate the functionality of our libraries' catalogs and Web sites so they can function as a coherent information system. Accomplishing this is difficult, because libraries' heroic accommodations to date are for the most part overlays and add-ons to the same operational and organizational models as before. We can do better, but to build truly coherent, usable and useful library systems that successfully integrate networked resources with our collections and services, we must be willing to transform ourselves, our methods, and our conventional organizational structures.

Toward New Models for Resource Description

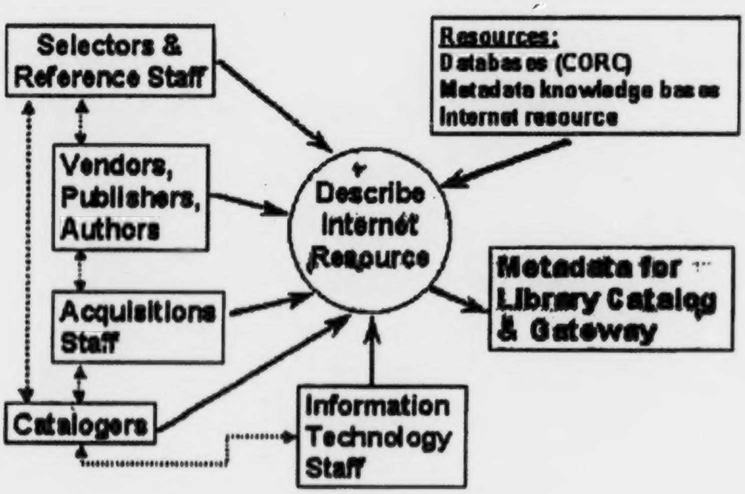
The call to fundamentally change organizational processes in order to achieve quantum leaps in organizational effectiveness is not new. In their pioneering book Reengineering the Corporation, Michael Hammer and James Champy (1993) call for "putting back together again the work that Adam Smith and Henry Ford broke into tiny pieces so many years ago." (20) They argue for "process teams" or "virtual teams" that obviate the need for hand-offs among functional departments and thereby dramatically speed up operations. These kind of teams are responsible not for a fragment of a process, but for an entire cross-functional process, such as handling a customer complaint, or completing the design of a new product or service. (21)

Along these same lines, I am proposing that building effective, coherent systems for the discovery and retrieval of library-selected resources will require us to brush aside conventional wisdom and start over. The first step is to put aside for the moment the assumptions we have made about the proper division of labor for producing resource descriptions. The second step is to step back from what we know about the appropriate level of description in a catalog and build systems that will allow library users to discover and navigate freely among resource descriptions and an array of heterogeneous collections. The third step is to rethink and expand our notions of standardization and controlled vocabulary.

How might we begin to reconceptualize organizational structures and processes for electronic resources? As demonstrated by Richardson (1999) in his modeling of the reference transaction, the tools of systems analysis provide for a graphic representation of inputs and outputs and a top-down perspective that can lead to new insights about a process. (22) The following section of this paper is an adaptation of Richardson's systems analysis approach; it attempts to model the electronic resource description process, and the players in it, as a system.

Figure 2 depicts the electronic resource description process as a widely distributed one, with many players inside and outside the library. It is intended to be an illustrative rather than comprehensive illustration of the process. Solid lines designate the flows of data in the system. The dotted line flows designate inquiries and responses exchanged among the various players (for the sake of clarity, some the inquiry/response lines have been omitted from the figure, but the intention is that inquiries and responses would flow between all players). The principal output of this system is metadata, or resource descriptions. It is assumed that the metadata provides a link to the resource itself.

Figure 2. Context of a Widely Distributed Electronic Resource Description Process


Selectors & Reference Staff
Resources:
Databases (CORC)
Metadata knowledge bases
Internet resource
Vendors, Publishers, Authors
Describe Internet Resource
Metadata for Library Catalog & Gateway
Acquisitions Staff
Catalogers
Information Technology Staff

In this proposed system, all functional groups in the library could conceivably contribute resource descriptions. Resource descriptions could also come from vendors, publishers, or authors themselves. Data will also flow in from resource description databases (e.g., CORC), from metadata knowledge bases (e.g., Dublin Core, EAD, and MARC/AACR2), and from the Internet resources themselves. The process ends with the output of metadata and its integration into the library information discovery and retrieval system. However, recognizing that the networked resource description process can be dynamic, the initial content of a resource description could be quite minimal, but it could then be modified and enriched over time.

As far as I know a fully-realized version of the system does not exist anywhere. But the proposed system can serve as a model for envisioning the context and data flows of future resource description processes. In addition, the model can help us respond creatively to the changes we face in libraries by freeing our minds from the past, stepping back from our mindsets, and allowing us to see what we need to do with fresh eyes. The following sections of this paper contain descriptions of various innovations at a number of libraries that have realized at least parts of the model of a widely distributed e-resource description process.

Selected Experimental Models

CORC at Brown and Cornell: Resource Descriptions from Collection Development, Reference and Cataloging

CORC is a cooperative effort to create a library-selected database of Web-based electronic resource descriptions. It is hosted by OCLC. (23) Prior to becoming production system in July 2000, CORC was a research initiative of several hundred libraries collaborating with the OCLC Office of Research. Some CORC participants have used the system as a framework for collaboration among different functional groups in their libraries. Two of these are Brown and Cornell.

The research phase of CORC became available at a time that librarians at Brown were discussing how to improve the information they provided to users to alert them to quality Web sites. They were interested in producing both Web-based bibliographies and cataloging records. Brown reference and collection development librarians used the fifteen elements of Dublin Core as the first step toward producing a MARC resource description for a Web resource. Catalogers then finished them in MARC and exported them to the local catalog. In addition, Brown subject specialists and reference librarians designed or customized pathfinders based on existing sites cataloged in the CORC database. In their article on CORC at Brown, Caldwell, Coulombe, Fark and Jackson (2001) express their satisfaction with the project outcomes, not the least of which was an enhancement of shared values among staff from different functional groups. They remarked, "for too long, it has been assumed that public and technical services cannot work closely together in a cataloging project because of differing agendas, missions, and skills. We at the Brown University Library have shown that it is possible and fruitful for both." (24)

The CORC project undertaken at Cornell was deliberately structured as a research project. We sought to take advantage of CORC's simultaneous support for Dublin Core and MARC as a breakthrough technology that would enable us to experiment, at low levels of risk and cost, with broadly distributed resource description.

For the duration of our experiment, we were a virtual team of three bibliographers, a reference librarian, two catalogers, and a project coordinator. The bibliographers and reference librarian selected the desired Internet resources and began the resource descriptions using Dublin Core. They supplied at least a title, URL, and a summary description. Next, the catalogers retrieved the records from the Cornell in-process file, finished them in MARC format, and then exported them to our local catalog and Gateway. While there was a hand-off of work from collection development/reference to cataloging, it was not the same as hand-offs between departments that are not organizationally integrated who in fact operate in functional "silos." Instead, the group functioned as a close-knit team, each member of which brought a different set of skills and perspectives to the work. Our workflow sought to put back together again a group that has been artificially separated by organization.

Cornell's experiences and results from the CORC project are reported elsewhere (Calhoun et al. 1999). (25) In summary, we found that the changes we made to the traditional workflow can ease and streamline the production of Internet resource descriptions. We found that distributed resource description is both feasible and beneficial, and that bibliographers and reference librarians can readily use Dublin Core to create preliminary records using CORC.

Another major plus was finding records already in CORC for most of the resources that were selected during the project. CORC is not only an extremely important advance in the library community's cooperative cataloging model, but it has the potential to expand the benefits of cooperation to new communities that need resource descriptions. Along the same lines, the team agreed that the most productive conversations about DC and MARC would assume that both have their place at Cornell. We should focus on how to forge a complementary relationship between the two standards, striving to optimize the strengths of each.

National Agricultural Library: Resource Descriptions from Authors

At the recent ALCTS Directors of Technical Services in Large Research Libraries meeting in Chicago, Sally Sinn of the National Agricultural Library (NAL) reported that scientists are creating and submitting resource descriptions of their work, and NAL librarians are working closely with them. There are two collaborations underway that are part of a redesign of NAL's RMIS (Research Management Information System) initiated by the Agricultural Research Service. One project is an effort that might be described as "indexing-in-publication" in which scientists submit descriptions for potential and completed publications (articles, book chapters, and conference papers) that is then standardized to match AGRICOLA citation format. In another project, NAL is developing a thesaurus of hierarchically arranged topics to describe the scientists' research projects for improved subject retrieval using standard vocabulary. (26)

Yale and the Record Set for EBSCO Academic Search Elite: Resource Descriptions from Vendors>

Late in 1998, the Program for Cooperative Cataloging's Standing Committee on Automation convened a Task Group on Journals in Aggregator Databases to (1) propose the content of vendor-supplied records for the full text journals in aggregator databases like ProQuest and (2) complete a demonstration project with an interested vendor. The initial task group completed its charge and issued a final report (Riemer and Calhoun 2000) (27), then was reconstituted for two more years to continue lobbying vendors to create record sets for their products and to pursue new areas of research.

In early 1999, EBSCO developers began collaborating with the task group to produce a record set for the approximately 1,100 titles of full text journals accessible from Academic Search Elite. The records, which are derived from CONSER records for the corresponding print journals, have been available for download by EBSCO customers at no charge since summer 1999. EBSCO periodically reissues the set to reflect additions, deletions and changes to their product.

Several libraries have acquired the record set from EBSCO and loaded them into their catalogs. One of the libraries was Yale. Matthew

Beacom, Catalog Librarian for Networked Information Resources at Yale, reported that the load generally went well.

Yale has also loaded the first set of updates to the initial load well as the records for EBSCO Business Source Premier titles. (28) By doing so, Yale has greatly enhanced its users' ability to discover via the catalog what full text resources are available to them, yet with a minimum of effort expended by its technical services and information technology departments.

The University of Tennessee-Knoxville and the Rochester Institute of Technology: Resource Descriptions from Information Technology Staff

A system of widely distributed resource description opens the door to broader participation of the library's information technology staff. Resource descriptions can be produced automatically, for example for the full text titles of e-journals in a vendor or publisher's database. Britten and others (2000) reported they harvested data about full text titles from vendors' Web sites and subsequently massaged them with Perl scripts and a utility called MarcMakr. (29) The end product was a set of MARC resource descriptions for the full text journals in several large aggregator databases for the catalog at the University of Tennessee at Knoxville.

At the Rochester Institute of Technology (RIT), a library wide task force was charged with finding an inexpensive solution to placing as many e-journals as possible under bibliographic control. Jiras (2000) reported that E-journal Web pages were becoming difficult to organize, providing full cataloging treatment for each title was too slow and labor intensive, and trying to keep up with added and cancelled titles and changes in holdings was a losing battle. The initiative that grew out of the task force's work led to a process in which library systems staff produced resource descriptions for aggregator e-journal titles by harvesting data from the vendor sites, massaging the data in several ways to produce MARC records, then loading the records into the catalog. (30)

Discussion

The library quadriad of collection development, public services, technical services and library systems is a persistent and highly successful organizational model. It is a functional division of labor that has the advantage of allowing specialization; in particular, catalogers' ability to focus on the catalog has produced millions of records for their own and the world's libraries in the past thirty years. Nevertheless, the functional division of labor has the disadvantages of fragmenting library processes, making cross-functional collaboration difficult, and discouraging "out of the box" thinking.

The incredible demand for Internet resources gives libraries strong incentives to reunite and intelligently coordinate the efforts of the individuals and groups that have always shared the work of resource description. In fact, an uneasy collaboration of cataloging, collection development, and reference librarians is already in evidence. There is already a two-pronged approach to the discovery and retrieval of electronic resources by users: catalog records and Web lists. But the two-pronged approach is too often uncoordinated and less than library users deserve. To do better, libraries must reintegrate the process of, and responsibility for, electronic resource description. Doing so is an essential first step in building a coherent, usable and useful library information discovery and retrieval system.

This paper proposes an electronic resource description process that could be very effective for making resource descriptions available more quickly, in greater numbers, and at less cost, assuming the process delivers metadata that is useful to readers. Lundgren and Simpson (1999) have explored the question of what is useful metadata. (31) I have touched on only a few of the experiments that are underway now to broaden and integrate participation in libraries' electronic resource description processes.

Looking Forward: Choosing Which Problems to Solve

Given the inevitable constraints on human, financial, and temporal resources, it is critical that librarians focus their energies on solving problems that will help their organizations in the future. As Buckland argued early in the decade, it is time to redefine the catalog, stop wasting effort on outdated models, and adopt a new bibliographic strategy. He urges librarians to think in terms of making use of all networked bibliographies and catalogs, not just local ones, and he lays out a set of basic functional requirements for a more universal approach to library collections. Similarly, in explicating their concept of "reference linking," Van de Sompel and Hochstenbach urge an evolution toward connecting all the available information, in order to come to a fully interlinked information environment. (32)

What might be some of the functional requirements of the information discovery and retrieval system proposed in this paper? One requirement would certainly be to deliver resource descriptions that are useful to readers. This would require us to answer the questions of what is a useful resource description, from a user's perspective, and how the needs of various user groups differ. As mentioned previously, Lundgren and Simpson have begun work on these questions, and I encourage others to take up where they left off.

Another requirement would be to support discovery and retrieval of a resource that is described at a full range of levels of granularity (e.g., citations and full text of articles, books, serials, sound recordings, images, the content of digital collections). A system that provides access to an array of information resources, both print and electronic, must provide contextual information and guidance to help users make sense of the results of their searches. (33)

A third system requirement would be a supporting infrastructure that brings order from the chaos inherent in this loose federation of data from many sources. Vellucci (1997) calls this a "metacatalog." It will contain resource descriptions in multiple metadata formats, created according to multiple standards, with name and subject headings created according to different communities' conventions, yet its infrastructure must present search results to searchers in a sensible way. On this point Vellucci says "the next generation metacatalogs should be able to access all relevant information seamlessly ... In order to accomplish this, each stakeholder community must ... concentrate on developing ways to layer, exchange and translate data within a loosely-coupled organizational system." (34)

The critical need for systems to be able to manage loosely federated data from many sources is far from unique to libraries, and it is not new. In 1998 the National Science Foundation, the Biological Resources Division of the U.S. Geological Survey, and an ALCTS task force hosted a Taxonomic Authority Files (TAF) Workshop. (35) The purpose was to bring together members of the biological sciences and library communities to explore the highly partitioned information environment in the biological sciences, to describe authority control in libraries, and to discuss the possibilities for managing widely distributed biological data sources to achieve consistency across shared concepts and names.

At the TAF workshop I presented an overview of authority control in libraries and concluded my talk with a call for a number of improvements to library authority control. Among these suggestions were to abandon the notion of a single, monolithic, all-encompassing global authority file in favor of a system of linked interoperable files; to deeply integrate authority data into end-user information systems (e.g., mapping a searcher's query into the vocabulary or naming conventions of the database being searched); and to better integrate the library community's authority control conventions with those of the abstracting and indexing community. I also noted that "taking any significant action would surely require a rethinking of the library community's current model of authority control." (36)

At the same workshop Stuart Nelson of the National Library of Medicine described the multi-thesaurus system called UMLS (Unified Medical Language System). (37) At the beginning of his talk Nelson used a Biblical analogy to illustrate the problems of a diverse, complex information system: "much of what the UMLS is approaching is ... [the problem] depicted in the story of the Tower of Babel." The purpose of UMLS is to retrieve and integrate information from patient records, databanks, bibliographic databases, full text sources, and elsewhere. An integral piece of the UMLS is a metathesaurus that includes data about naming conventions in a variety of different systems using a variety of controlled vocabularies.

These are the kinds of problems that librarians who are building the next generation of information discovery and retrieval systems will need to grapple with and solve.

Conclusion

Libraries appear willing to experiment, but I anticipate many obstacles to the full deployment of a widely distributed electronic resource description process. I believe the principal obstacles will not be technical or operational, but organizational and attitudinal. Many librarians are deeply vested in existing processes and organizational structures. Not only that, the existing processes and structures still function well for most items that are added to the collections. Finally, because we are and will be in a transitional state for some time, librarians must strike an appropriate balance between their everyday work and new ways of doing things.

Perhaps in the near term, then, it would be more practical for libraries to avoid radical restructurings and instead make liberal and frequent use of virtual teams. These are cross-functional groups that exist alongside (and sometimes outside) the formal organizational structure. For example, for the purpose of introducing a new electronic resource description process, or providing ongoing support for one, the members of the virtual team would share authority and accountability. At the same time, the members would continue to report to different individuals and departments in the library hierarchy.

Beyond the redesign of library workflows, building the coherent, usable and useful information discovery and retrieval system I have proposed will require determination, perseverance, and skills from all walks of librarianship. I am convinced that technical services librarians have a great deal to bring to the table, provided they tap into their creativity and apply their significant knowledge of bibliographic control to the new information space in which libraries operate.

---

1. Younger, Jennifer A. and D. Kaye Gapen. 1990. Technical services organization: where we have been and where we are going. In Technical services today and tomorrow, ed. Michael Gorman. 171-83. Englewood CO: Libraries Unlimited.

2. Beile, Penny M. and Megan M. Adams. 2000. Other duties as assigned: emerging trends in the academic library job market. College & research libraries 61, no. 4: 336-47.

3. Stueart, Robert D. and Barbara B. Moran. 1993. Library and information center management. 4th ed. Englewood CO: Libraries Unlimited, 80-1.

4. Younger and Gapen, Technical services organization, 176.

5. Boissonnas, Christian M. 2001. Technical services: the other reader service. portal: libraries and the academy 1. In press.

6. Thomas, Sarah E. and Jennifer A. Younger. 1993. Cooperative cataloging: a vision for the future. Cataloging & classification quarterly 17, nos. 3, 4: 257.

7. Ferguson, Chris. 2000. "Shaking the conceptual foundations," too: integrating research and technology support for the next generation of information service. College & research libraries 61, no. 4: 300-11. The title within the article title refers to Jerry Campbell's 1992 article in Reference services review. While Ferguson is discussing the dramatic changes in reference services and calling for dramatic new approaches, the parallels with the history and current environment in technical services are striking. I particularly appreciated Ferguson's insight into the difference between "layering" new services upon old models, versus creating new service models, as well as his use of the phrase "heroic accommodations," which I have borrowed for this article.

8. Buckland, Michael. 1992. Bibliographic access reconsidered. Ch. 4 of Redesigning library services: a manifesto. 24-41. Chicago: American Library Association.

9. For the purpose of this paper, resource description is defined broadly to include the creation of any surrogate of an item whose purpose is to facilitate library users' discovery and retrieval of the item.

10. Buckland, Bibliographic access reconsidered, 29.

11. Lubans, John. 1999. Students & the Internet: spring 1999 survey (study 3). Available at: http://www.lib.duke.edu/lubans/docs/study3.html. Accessed: August 11, 2000.

12. Calhoun, Karen and Zsuzsa Koltay. 1999. Library gateway focus groups report, January 1999. Available at: http://www.library.cornell.edu/staffweb/GateEval/contents.html. Accessed: August 11, 2000.

13. Buckland, Redesigning library services, 32-9.

14. Van de Sompel, Herbert and Patrick Hochstenbach. 1999. Reference linking in a hybrid library environment, parts 1 and 2. D-Lib magazine 5, no. 4 (April). Available at: http://www.dlib.org/dlib/april99/04contents.html. Accessed: August 11, 2000.

15. Ercegovac, Zorana. 1997. The interpretations of library use in the age of digital libraries: virtualizing the name. Library & information science research 19, no. 1: 35-51. Table 1 is an adaptation of Ercegovac's table "Toward digital libraries" on page 42.

16. Buckland, Redesigning library services, 32-

17. Neuhauser, Peg. 1988. Characteristics of organizational tribes. Ch. 2 of Tribal warfare in organizations. 15. Cambridge MA: Ballinger Publishing.

18. I chose seventeen titles to look for--ten online databases and seven full text journals. The list below provides the titles and the sources from which they are available. I selected them because (1) they are commonly licensed by large ARLs and thus likely to be accessible to their users; and (2) they present resource description challenges of different types and levels (simple to complex). For example, Callaloo is available in JSTOR, a full text collection with a stable list of titles that are well maintained over time, while American Heritage is part of several large, amorphous vendor aggregations with shifting sets of titles. The seven libraries were Harvard, Yale, UCLA, the University of Illinois at Urbana, the University of Michigan, Columbia, and Cornell.

Titles Examined	Source(s)
ABI Inform	Ovid, OCLC, Proquest
Avery index	RLG
Arts & humanities citation index	ISI Web of Science
Congressional Universe	CIS
Dissertation abstracts	UMI
ERIC	ERIC, Ovid, OCLC
INSPEC	IEE, Ovid
Academic Universe	Lexis Nexis
UnCover	CARL
WorldCat	OCLC
Harvard business review	Full text in Ovid ABI Inform, Proquest ABI Inform, EBSCO Academic Search Elite, etc.
Time	Full text in Ovid ABI Inform, Proquest ABI Inform, EBSCO Academic Search Elite, OCLC Periodical Abstracts, etc.
Callaloo	Full text in JSTOR, Project Muse, OCLC ECO, etc.
American Heritage	Full text in EBSCO Academic Search Elite, OCLC Periodical Abstracts, Proquest Periodical Abstracts, etc.
SIAM journal on applied mathematics	Full text in JSTOR, SIAM Journals Online, etc.
Wall Street Journal	Full text in Proquest Periodical Abstracts Research II, OCLC
Algorithmica	Full text in Springer Link, EBSCO Online, etc.

19. There were cases in which the library did not license the title in question, but these cases were rare. Usually the missing title was licensed, just not findable via the library's catalog or Web site.

20. Hammer, Michael and James Champy. 1993. The new world of work. Ch. 4 of Reengineering the corporation: a manifesto for business revolution. 65. New York: HarperBusiness.

21. Ibid., 66-7.

22. Richardson, John V. Jr. 1999. Understanding the reference transaction: a systems analysis perspective. College & research libraries 60, no. 3: 211-22.

23. For more information about CORC visit http://www.oclc.org/oclc/corc/index.htm. Accessed: August 11, 2000.

24. Caldwell, Ann Dominique Coulombe, Ronald Fark, and Michael Jackson. 2000. Never the twain shall meet? Collaboration between catalogers and reference librarians in the OCLC CORC project at Brown University. Journal of Internet cataloging 4, no. 1. In press.

25. Calhoun, Karen, et al. 1999. CORC at Cornell project: final report. Available at: http://campusgw.library.cornell.edu/corc/. Accessed: August 11, 2000.

26. Sinn, Sally. 2000. Reported in the minutes of the ALCTS Directors of Technical Services in Large Research Libraries meeting, July 7, 2000, and confirmed in e-mail exchange with the author, August 8, 2000.

27. Riemer, John and Karen Calhoun. 2000. PCC Standing Committee on Automation (SCA) Task Group on Journals in Aggregator Databases: final report, January 2000. Available at: http://lcweb.loc.gov/catdir/pcc/aggfinal.html. Accessed: August 11, 2000.

28. Beacom, Matthew. 2000. E-mail exchange with the author, August 8, 2000.

29. Britten, William A., et al. 2000. Access to periodicals holdings information: creating links between databases and the library catalog. Library collections, acquisitions and technical services 24, no. 1. In press.

30. Jiras, Jonathan. 2000. Access to e-journals at RIT. Presentation at Partners in Information and Innovation meeting, February 2, 2000, at Rensselaer Polytechnic Institute, Troy, NY.

31. Lundgren, Jimmie and Betsy Simpson. 1999. Looking through users' eyes: what do graduate students need to know about Internet resources via the library catalog? Journal of Internet cataloging 1, no. 4: 31-44.

32. Van de Sompel and Hochstenbach, Reference linking, part 1.

33. For example, in a fully interlinked information environment, a user's search for "George Washington" could retrieve images, correspondence and other primary source materials, books, articles in journals, audiovisual materials, etc. How can we provide enough information about where the user's "hits" are coming from to allow him or her to make sense of what has been retrieved and to navigate to what is wanted?

34. Vellucci, Sherry V. 1997. Options for organizing electronic resources: the coexistence of metadata. Bulletin of the American Society for Information Science 24, no. 1 (Oct./Nov.).

35. A description of the TAF workshop and the participants' papers may be found at http://research.calacademy.org/taf/proceedings/Proceedings.html. Accessed: August 11, 2000.

36. Calhoun, Karen. 1998. A bird's eye view of authority control in cataloging. Paper presented at TAF Workshop, June 22-23, 1998, in Washington DC. Available at: http://research.calacademy.org/taf/proceedings/Calhoun.html. Accessed: August 11, 2000.

37. Nelson, Stuart. 1998. The Unified Medical Language System: applicable experiences and observations. Paper presented at TAF Workshop, June 22-23, 1998, in Washington DC. Available at: http://research.calacademy.org/taf/proceedings/nelson/index.htm. Accessed: August 11, 2000.

Library of Congress
December 14, 2000
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Metadata, Cataloging, Digitization and Retrieval: Who's Doing What to Whom: The Colorado Digitization Project Experience

About the presenters:

Liz Bishoff

Project Director
Colorado Digitization Project

Liz Bishoff is currently the Project Director of the Colorado Digitization Project. The project, a collaborative among Colorado's libraries, museums, archives and historical societies, is developing a virtual collection of Colorado's unique resources and special collections. The project, funded by IMLS, the Colorado State Library and the Colorado Regional Library Systems has awarded $180,000 in grants to 40 institutions involved in 30 projects. The projects are creating digital images on a range of topics from the University of Colorado Boulders historic sheet music collection to the National Mining Museum and Hall of Fame historic photograph collection, to the Boulder History Museum's historic costume collection. A total of 50,000 images will be created. To provide enhanced access to these resources, the CDP is developing a union catalog of metadata using a variety of metadata formats. The project has developed standards and guidelines for metadata and scanning, a website that brings together approximately 40 existing digitization projects, and a training program for participants. All is accessible via the website http://coloradodigital.coalliance.org.

Liz is the owner of The Bishoff Group, a management consulting organization specializing in library and library related organizations.

Prior to her current position, Liz was Vice President, Member Services at OCLC. Her responsibilities include management of OCLC relationships with external organizations, including the national libraries, professional library-

Cataloging Directorate Home Page

Library of Congress Home Page

related organizations and government relations, OCLC Users Council, and Library Member Relations. Liz was actively involved in many national cooperative cataloging programs, including CONSER and was a founding members of the Program for Cooperative Cataloging. Prior to this position, Liz was Director of the Online Union Catalog Product Management Division, which included strategic planning and product management for OCLC PRISM Cataloging, Interlibrary Loan, and Union List systems.

Liz is the immediate past-President of the American Library Association, Association for Library Collections and Technical Services. She is currently the ALA Treasurer and a member of ALA Board. Liz has more than 30 years of work in the cataloging, including membership on the Decimal Classification Editorial Policy Committee, a member of the ALCTS Subject Analysis Committee, and a member of the Catalog Code Revision Committee. In addition to her involvement in ALCTS, she has also held committee appointments in the Public Library Association and the LAMA.

Liz has extensive experience in public libraries. She was the principal librarian for Support Services at Pasadena (California) Public Library, with responsibility for management of the technical services, circulation and automated services. Liz has been a public library director, school media specialist, and cataloger in her 30 year library career. She has taught in the graduate library programs at Rosary College and Emporia.

Liz holds an MLS from Rosary College, and has post-graduate work in public administration at Roosevelt University.

William A. Garrison

Head of Cataloging
University of Colorado, Boulder
and Member of the CDP Metadata Working Group

Bill received his MLS from Rosary College (now Dominican University) in 1979 and has been involved in cataloging or cataloging related activities for his entire career. He is currently Head of Cataloging at the University of Colorado at Boulder and has previously held positions at Stanford University and Northwestern University.

He has been active in the Program for Cooperative Cataloging (PCC) serving on the Standing Committee on Standards and on the BIBCO Operations Committee. In addition, he serves as a trainer for the PCC for both NACO and BIBCO. He has conducted NACO training at the National Library of New Zealand, the University of California at Los Angeles, the University of New Mexico, the Nevada State Library, Trinity University (San Antonio), the University of Wyoming, and the University of Kansas. His BIBCO training includes the University of Oregon, Texas A&M University, the University of New Mexico, and Oklahoma State University.

Bill has also been active in ALA/ALCTS and is currently the Chair of the ALCTS Cataloging and Classification Section. Previously he has served on the ALCTS Membership Committee, the ALCTS Fundraising Committee, the ALCTS Leadership Development Committee, ALCTS/CCS Subject Analysis Committee, and ALCTS/CCS Policy and Research Committee. He has also given many papers at ALA conferences and has published in various professional journals.

In Colorado, Bill served as Chair of the Cataloging and Reference Task Force that designed and implemented Prospector, a union catalog for 16 institutions in Colorado, and has worked on the Metadata Working Group of the Colorado Digitization Project (CDP) since the CDP's inception. He has taught metadata workshops for the CDP and worked on the standards devised by the CDP for project participants. In addition, he has served as a web-mentor for students at Dominican University.

Full text of paper is available

Summary:

The Colorado Digitization Project, a collaborative of Colorado's archives, historical societies, libraries and museums has undertaken an initiative to increase user access to the special collections and unique resources held by these institutions via digitization and distribution via the Internet. When exploring the holdings of these nearly 350 institutions, we find that there is significant overlap in holdings, not overlap of individual items, but content overlap. The goal of the CDP is to find ways to bring together the resources held by widely dispersed cultural heritage institutions into one virtual collection. The CDP website will provide one stop shopping for the residents of Colorado and beyond.

Over the last 24 months, the project has had to address a range of issues to realize our goal of increased access. Many of these issues have emerged as a result of the multi-cultural heritage institution types participating in the project, including the lack of a common mission or vision, different audience expectations, insufficient knowledge base on the range of issues related to digitization, the lack of a common set of metadata standards, both for the descriptive components and the subject analysis, urgent need by users to locate this widely distributed content, barriers presented by current web searching, and unfamiliarity working across cultural heritage institution types.

Through the first year (1998-1999) the project began by building the collaborative, exposing participants to the different needs and issues of the partner organizations. All agreed that to realize the goals of the project, standards, particularly metadata standards had to move to the top of the list. Second, there was the recognition that we couldn't mandate a single metadata standard, as many of the institutions had systems in place to support their internal metadata needs. Third we realized that it would be

years before the web searching would be sophisticated enough to retrieve the level of information from a decentralized set of images.

Our first decision was to develop a union catalog of metadata, as a near term solution to the information identification issues. To support that union catalog, and accommodate local preferences, we developed a set of metadata guidelines (descriptive, functional and administrative) that doesn't require the adoption of one standard, such as MARC or EAD or DC. Rather we established a set of core elements derived from the Dublin Core elements, which when loaded into the OCLC SiteSearch software would support cross database searching.

Library of Congress
August 16, 2000
Comments: lcweb@loc.gov

Metadata, Cataloging, Digitization and Retrieval: Who's Doing What to Whom: The Colorado Digitization Project Experience

For the Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the challenges of networked resources and the web

Submitted by
Liz Bishoff, Project Director
Colorado Digitization Project
And
William A. Garrison, Head of Cataloging
University of Colorado, Boulder
November 2000

Final version

In the last five years there has been significant growth in museum/library collaboration, in part due to the Institute of Museum and Library Services national leadership program and in part due to the growing realization that both libraries and museums are holders of collections that represent our rich and diverse culture heritage.

Museum/library collaboration isn't just occurring in the United States. In 1999, the European Commission's Information Society Directorate General appointed a working group to develop a research framework for archives, libraries and museums that would support their work in the networked environment. The primary purpose of the research framework is to support access to resources available on the Internet. The document notes that the framework is based on the assumption, "...that libraries,

430

archives and museums have shared research interests... can identify several broad goals that underpin these and encourage collaborative activity.... " [1]. The goals are:

- To release the value of Europe's scientific, industrial and cultural heritage in creative use by its citizens.
- To engage with the cultural identities and aspirations of Europe and its peoples.
- To develop practices appropriate to upholding the values and purposes of library, archival and museum traditions in a digital environment.
- To explore what it means to develop virtual civic presence.
- To explore sustainable economic models which support both development and continued equitable access to the cultural heritage. [2]

While these institutions share similar goals and missions, there is no common vocabulary, no common policies on access and use by the public, no common term for this group of institutions, and no common standards to support the goal of access.

The report summaries that these institutions:

- Organize the European cultural and intellectual record
- Contain the memory of peoples, communities, institutions and individuals, the scientific and cultural heritage, and the products through time of our imagination...
- They join our ancestors and are our legacy to the future generations.
- Support the creation of the heritage of the future. [3]

Within this common vision, each of the communities addresses the goals within their own curatorial traditions and organizational contexts, and specific national or administrative framework. "The recognition that common interests converge on the Internet, driven by the desire to release the value of their collections...that support creative use by as many users as possible." [4] The participating institutions understand that users desire increased access to the intellectual and cultural materials in a flexible manner, without concern for who owns the resource. "To support this need, they recognize the need for services that provide unified routes into their deep collective resources...." [5]. At the same time these institutions are all developing their own approaches for organization and access to their resources. They may be working with subject based peer institutions across the continent or internationally to develop versions of Dublin Core (DC) or the Encoded Archival Description (EAD); or they maybe working within their type of organization to develop Visual Resources Association (VRA) description for visual resources. There is little evidence of work across institutions of different types at the implementation stage.

Assuming that U.S. museums, libraries and archives share the same goals and vision as our European colleagues, then the issues discussed at the 'Bicentennial Conference on Bibliographic Control for the New Millennium' must be discussed within a community that involves our museum and archive colleagues. For as the EC paper notes, without providing our common users with a means of identifying the unique resources and special collections, the mission of access to our heritage will be severely restricted. Several papers, including that by Caroline Arms, touch on the issues related to the collaboration of many institutions on the American Memory Project [6].

This paper will focus on the specific experiences of the Colorado Digitization Project (CDP) related to accessing a diverse set of primary resources held by many different cultural heritage institutions. The paper will address issues that arise from different cataloging and metadata standards and diverse user populations and needs. The biggest challenge for the CDP is to bring metadata from the various institutions together in a single union catalog and to present the user with retrieval of digital objects stored in a distributed network environment.

Description of the project:

The Colorado Digitization Project begun in the fall of 1998, is a collaborative initiative involving Colorado's archives, historical societies, libraries, and museums. The CDP's goal is to create a virtual digital collection of resources that provide the people of Colorado access to the rich historical, scientific and cultural resources of the state. Project participants will be able to contribute content that has been reformatted into digital format, as well as the born digital. The virtual collection will include such resources as letters, diaries, government documents, manuscripts, music scores, digital versions of exhibits, artifacts, oral histories, and maps.

Initial funding from the Colorado State Library supported the development of the collaborative, identification of ongoing and planned digitization initiatives, development of best practices for digitization projects, a small pilot project and identification of future funding options. For fiscal year 1999-2001, the CDP was awarded a two-year $499,999 grant from the Institute on Museum and Library Services and a second LSTA grant of $107,000. In addition, the Regional Library Systems of Colorado awarded the CDP a $36,000 grant. The grant funds supported the expansion of the project to include:

- Establishment of 5 regional scan centers
- Training for Colorado archivists, librarians and curators
- Creation of a union catalog of metadata
- Financial support for 20-25 collaborative digitization initiatives
- Research on two key issues
- Creation of 50,000 new digital images

The CDP Strategic Plan for 1999-2002 http://coloradodigital.coalliance.org/about.html, establishes the project goals:

- To create an open, distributed, publicly accessible digital library that documents key information for the residents of Colorado,
- To expand the collaborative structure among the state's libraries, museums, archives and historical societies to coordinate and guide the implementation of a virtual digital collection,

- To establish criteria and standards to guide the selection of materials for inclusion in the digital library,
- To demonstrate the value of libraries/museums in the emerging electronic information environment and their important contribution to the state's development,
- To assist libraries, archives, historical societies, and museums in the digitizing of materials and managing digital projects through training programs and consultation,
- To emphasize the content and rich resources held by Colorado archives, historical societies, libraries and museums, and
- To work with the Colorado K-12 environment to incorporate digital objects that assist teachers, parents and students in meeting the Colorado history standards.

To implement the plans, the CDP has a variety of working groups with membership from different constituent groups. These groups were responsible for developing best practices for metadata and scanning. The CDP website (http://coloradodigital.coalliance.org) introduced in January 1999, provides access to resources and information about the project, the best practices on metadata and scanning, links to digital resources and information on legal issues. As of summer, 2000, the website links to more than 40 digital collections available in Colorado. That number will be doubled as the funded projects come online.

As part of the IMLS grant, the CDP will conduct two research projects, the first focusing on the impact that digital images available via the internet will have on museum attendance, and the second a project researching user satisfaction with two approaches for providing access to digital objects, the exhibit/interpretative approach vs. the catalog/database approach.

***Environment for standards application in a cross-cultural heritage institution group*:**

In order to meet the objective of increased access to digital collections, the first effort undertaken by the CDP was identifying the approaches used among the existing projects to provide access to their collections. Among the initial 15 projects, there were 8 libraries and 7 non-library participants. Among those there were a range of approaches to providing access to the collections. Several provided access through their local library system and MARC records. Several presented exhibits with an additional database to search for individual digital objects. Many offered only exhibits, while two provided access via a locally developed database. One university library offered collection level MARC records linked to HTML finding aids, finally linking to images. Clearly even at this early stage, there was no dominant approach and therefore little possibility of a single standard or a single search engine. This is due, in part, to the lack of a dominant standard, the early stage of development of systems supporting access to digital objects through the new standards, and in part because of a lack of a funded mandate that would provide for a single system or approach. Additionally when a web search was undertaken these sites frequently weren't located, as they were several layers down on the host website. Where a database supported searching of specific images, the images weren't located, as the web engines cannot search a subsequent database.

Outside the library community, these organizations either used a specialized standard for description and specialized thesauri or taxonomies or they created their own with some providing no metatags at all. In the library and archival community there was use of collection level description and item level description. None of the current or planned projects had adopted the Encoded Archival Description, Dublin Core, Text Encoding Initiative or any of their derivative standards.

Like the European Community, the CDP found that there was a lack of common vocabulary, lack of common software, and a lack of standards that would support interoperability.

The CDP and standards:

It is within this environment that the Metadata Working Group began its work. In addition to understanding the approaches taken by current and planned projects, the group reviewed current and emerging standards, including EAD, MARC, Government Information Locator Service, DC, VRA, etc., for common elements. As web searching would not provide the desired access, and a single centralized metadata and image system would not be politically or financial feasible, the working group recommended the development of a union catalog of metadata to provide a desired level of access, hoping that future developments in web searching would negate the long term need for the union catalog. The guidelines are intended to promote best practices and consistency in the creation of metadata records across the different cultural heritage institutions and skill levels, while enhancing online search and retrieval accuracy, improve resource discovery capabilities and facilitate and ensure interoperability. To achieve this objective, institutions must create metadata or cataloging data at a sufficient level to support the identification and access needs.

The metadata standard chosen by an institution depends on a variety of factors. These factors include the type of materials that are being described and digitized, the purpose of the digitization project (access or preservation or both), the user community, the knowledge and expertise of project staff, and the technical infrastructure of the institution. The level of detail for a resource also varies from institution to institution. Information may be proprietary or confidential and may not be distributed or accessible on systems open to public access. Agreement on inclusion of such administrative information is unlikely. As a result, the Metadata Working Group determined that information of this type would not be retained in the union catalog record.

The CDP Core Elements:

Based on the analysis of the metadata standards, the working group recommended adoption of the Dublin Core/XML metadata for the union catalog. Rather than adopting a specific communication form such as MARC or EAD, the working group developed a minimum set of elements that must be included in a cataloging or metadata record based on the fifteen Dublin Core elements. The working group recognized that additional elements might be required for particular formats and has accommodated this in its recommendations.

The recommendations of the group for the "core" and "full" record in Dublin Core are as follows:

Mandatory Elements:	Optional (Desirable) Elements:
Title	Contributor
Creator	Publisher
Subject	Relation
Description	Type
Identifier	Source
Date Digital	Language
Date Original	Coverage
Format View	Rights
Subject: Classification number	
Identifier: Owning Institution	

The "mandatory" or "core" elements were designed along the same guidelines as the core records for the Program for Cooperative Cataloging were developed. In addition, the working group recommended that a "qualified" Dublin Core be implemented. This record employs modifiers and schemes for each element as appropriate. For example, a recommendation that subject terms from a recognized thesaurus be used has been made. The CDP Metadata Guidelines http://coloradodigital.coalliance.org/guides provide links to all publicly accessible subject heading lists and thesauri.

Each element of the Dublin Core has been defined. For example, the subject element has a web page as follows:

Subject

Label: Subject

Definition: Topic of the digital resources. Typically, subject will be expressed as keywords or phrases that describe the subject content of the resource, or terms related to significant associations of people, places, and events, or other contextual information.

Mandatory: Yes

Repeatable: Yes

Scheme: Use established thesaurus: Library of Congress Subject Headings (LCSH), Art and Architecture Thesaurus (AAT), Thesaurus for Graphic Materials (TGM), Medical

Subject Headings (MESH), ICONCLASS, etc.

Input guidelines:

1. *Prefer use of most significant or unique words, with more general words used as necessary*
2. *Subjects may come from the title or description field, or elsewhere in the resource*
3. *If the subject is a person or organization, enter as outlined under Creator*

Examples of subject terms/descriptors are also provided.

***Issues with Dublin Core*:**

Adopting the Dublin Core framework at this early stage is risky; however it is likely to be the best option for integrating records using a variety of international best practices/standards. Adopting Dublin Core in 2000 is like adopting MARC in 1970. Early adopters of MARC recognized that there would be changes to MARC, that the systems would be available to support it, etc. We are facing similar issue in 2000 with Dublin Core. As the project was focusing on metadata for digital objects vs. websites, significant interpretation was required. Most problematic for the working group was the handling of the date for the original object, which was needed to qualify searches. Using the Source field for this information would negate the possibility of qualifying searches by date. After many discussions, the group decided to add an additional date field to accommodate the original date. The other aspect that caused the group difficulty was accommodating the functional metadata relating to the digital object. Again after much discussion, the group decided to use the Format field for both the requirements for use of materials and a second Format field for the requirements for creation of the resource. Lastly the group added a field for holding institution, allowing the user to limit searches by the owing institution.

As noted in other papers, software supporting both the creation and use of Dublin Core based records is slow to develop and implementation is unsettled due to the evolving nature of the standard. The advantage of adopting Dublin Core is that many specialized communities, archives, libraries and museums are creating Dublin Core based derivatives for their communities.

What do you describe?

Not unexpectedly, the issue of cataloging the original versus cataloging the digital object has arisen, regardless of whether the owning institution is a museum or library. Some institutions catalog the original item, providing a link to the digital image/object. This practice, in most instances, does not preserve or record any of the details of the digital object (e.g., scanning equipment, resolution, rights management, etc.). In many of these cases, it is a financial decision. The cataloging already exists for the original and the most cost effective approach is to provide access to the digital version by adding a URL or other linking identifier. In many instances the digital object is considered secondary to the original, so where the original item is not cataloged, cataloging for the original is preferred, with the URL linkage to the digital. The public service and reference librarians have also expressed concern for multiple records for the same item. This discussion is not dissimilar to the multiple version discussions the library community has had for more than two decades.

Some institutions catalog at the collection level and not at the item level, others catalog at the item level only, and others catalog both at the collection and item level. Within one institution all three approaches have been taken. Those that provide access to the digital object through a collection level record, generally have finding aids. As with original cataloging, the existing finding aid is converted to HTML rather than to another format. As finding aids focus on the hierarchical relationship of the items within the collection, there is little subject rich terminology for the item level materials, limiting access to the individual resources in the collection. In response, institutions are expanding the subject terms for the collection level cataloging. With the future hope of full text indexing of the resources in the collection, enhanced retrieval is a possibility, but until then the only other option is providing the enhanced subject terms in the finding aids themselves.

To accommodate the different approaches and different standards, the CDP licensed the OCLC SiteSearch software to build its union catalog for accessing the digital collections in Colorado. The SiteSearch software allows CDP participants to batch load records into the system and supports online record creation. The CDP is working with OCLC on enhancements to the software, as there are currently limitations on the variety of formats handled. It is anticipated that SiteSearch, as implemented by CDP, will enable participants to contribute records in a variety of formats. A loader profile has been developed for the CDP participants. Initially records may be batchloaded in either MARC format or SGML/XML. The SGML/XML capability will be used to load locally developed databases, as well as commercial databases supporting the museum and historical society communities. The capability to load records in Encoded Archival Description (EAD) as well as records in other formats (e.g., VRA) is being explored with OCLC. Initially the CDP had planned to use the SiteSearch record builder capability allowing input in either Dublin Core or MARC, but due to time constraints in implementation, the CDP will offer a locally developed search intake mechanism for online input. These online records, built with a Dublin Core template, the MARC records loaded from library local systems, and the SGML/XML loaded records will create a single union catalog. All records will be converted to the CDP defined Dublin Core elements.

Among the features that CDP hopes to have incorporated into SiteSearch in the future are the ability to load records in formats other than MARC and SGML/XML, the ability to export records, an authority control feature/system, and an improved online entry and maintenance system. While SiteSearch has been specifically designed as a library "system", CDP is expanding the system to meet the needs of the varied cultural heritage institutions involved in this collaborative venture.

Subject terminology:

A wide range of issues exists in the area of subject retrieval in the CDP Union Catalog. The mix of cultural heritage institutions resulted in many specialized institutions, for example the Florissant Fossil

Beds National Monument and their collection of 6000 unique fossils, or the Crow Canyon Archaeological Center and the large collection of archaeological materials or the Boulder History Museum and their more than 4000 costumes and accessories. The first two use taxonomies from their specialized fields, while the third uses *The Revised Nomenclature for Museum Cataloging, A Revised and Expanded Version of Robert G. Chenhall's System for Classifying Man-Made Objects* by James R. Blackaby, Patricia Greeno, and the Nomenclature Committee. Published by American Association for State and Local History, 1988. At the same time some of the smaller or more general collections will contain these type of resources or subjects, but use a more generalized subject heading list such as the *Library of Congress Subject Heading List*. The CDP Union Catalog will provide access to this entire range of terms without an authority control system. As a result unless the user knows both the general and specialized taxonomy, retrieval will be limited to the term input. To address this situation, the project is testing the use of Dewey Decimal Classification numbers that will be assigned to each record, allowing the linkage of general terms and highly specialized terms within a browse feature. When using the keyword or advanced search capabilities the users will retrieve only the term/terms entered, a common approach for both museums and libraries.

The project is addressing one area of authority control, terms for Colorado geographic names and subjects. In order to assure some level of consistency in terminology, the CDP has developed a list of Colorado terms that a user can search from the SiteSearch web. The list can be searched by specific term or through a browse function. The list is being created by extracting headings from the Prospector database, the database reflecting the collections of Colorado's major public and academic research libraries, as well as the community colleges and four year schools. The Metadata working group has begun exploring the idea of turning this list into a real thesaurus and/or a full authority file. The later would be approach through statewide NACO/SACO project creating name headings and subject headings to be added to the *Library of Congress Name Authority File* and *Library of Congress Subject Heading List*.

What needs to be addressed in the shared cultural heritage environment?

Shared development: In order to reach commonality in standards and address the interoperability issues, participants from across the range of institution types need to be at the table at the start of the discussions. Libraries cannot determine the standards and assume that museums, archives and other cultural heritage institutions will adopt them.

Standards: The key to participation of a wide range of institutions lies in the ability to allow the metadata creators to use multiple standards while attempting to ensure that there is agreement between the various standards for some commonality in the access points provided. This will clearly call for the cultural heritage institutions (including libraries) to have discussions related to access and interoperability issues. Assuming that some commonality among/between the various standards can be reached, there will clearly be an impact on the search engines used to access these resources.

Interoperability: Many projects state that they have as an objective the interoperability of the systems;

however, when queried, interoperability means adoption of a single set of standards and use of a single system or adoption of one vendor's software. At this time the predominant communication format for libraries, MARC, doesn't support the descriptive elements required by museums and archives. The same is true for museum based software, it doesn't meet the standards of libraries. With the development of XML and Dublin Core there is some hope that a system meeting the different needs may be accommodated.

Resource discovery services: With the development of the OCLC CORC service, we have the first opportunity to build a resource discovery service that supports standards (Dublin Core) that have possible use by different cultural heritage institutions. Unfortunately OCLC services are library-centric. Adoption of CORC by non-libraries will be not come easily as the system development did not include non-library representation and input.

Cataloging issues: Cataloging differences also pose some challenges. The cataloging of three-dimensional objects provides a good example. The museum community typically does not assign titles to such objects whereas libraries routinely supply titles to objects or items that lack them. The question arises: does it make a difference if there isn't a title supplied? How is retrieval affected? Another example occurs with the level of specificity applied in subject analysis. A very small historical society may not need the same level of specificity in the description of its materials, as does a large historical society, library or museum. What impact will different levels of subject analysis and specificity have on retrieval?

Authority control: Our discussions of authority control innovation must also include use of taxonomies as well as thesauri and subject heading lists. Barbara Tillett's suggestion of a single integrated authority record sounds appealing, however complicated [7]. The subject "field" as defined in Dublin Core with the appropriate scheme qualifiers almost presumes an ability for a system/search engine to perform cross-vocabulary searching. This certainly also poses a whole different set of challenges.

Will we succeed?

We expect to succeed. To do that, the best practices will have to become standards and the standards will have to continue to evolve, much as MARC has, and most important, the standards will have to be adopted. It is only when the standards are adopted that systems will be developed to support the widespread use. For us to achieve the vision of providing our citizens with the broadest possible access to the cultural resources of our peoples, we will need to develop standards and systems that have broad-based adoption across the different cultural heritage communities. To do that, we have to sit down at the table together. The people at today's conference have the opportunity to take a leadership role in calling together the cultural heritage institutions of the United States to begin working on the issue of how to increase access to our collective digital resources.

Notes

1-5 "Scientific, Industrial and Cultural heritage: a shared approach; a research framework for digital libraries, museums and archives," Ariadne, Issue 22.
6. Arms, Caroline, "Some Observations on Metadata and Digital Libraries," Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000.
7. Tillett, Barbara, "Authority Control on the Web," Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000. (http://www.loc.gov)

Library of Congress
November 7, 2000
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Michael Kaplan

Director, Product Management
Ex Libris (USA), Inc.
1919 North Sheffield
Chicago, IL 60614-5018

Exploring Partnerships: What Can Producers and Vendors Provide?

About the presenter:

Michael Kaplan is currently Director of Product Management for Ex Libris (USA),. Inc. Previously, he spent two years as Associate Dean of Libraries & Director of Technical Services at the Indian University Libraries in Bloomington. From 1977 to 1998, he worked in various technical services positions at Harvard University. For much of the 1990s he was actively involved with the Library of Congress, where he participated in the Seminar on Copy Cataloging in 1992, and since 1993 with the Program for Cooperative Cataloging (PCC). He served the PCC as chair of its Standing Committee on Automation from 1993-1998. In 1998-1999, he was a member of its Policy Committee and was chair of the PCC in 1999-2000.

Dr. Kaplan has been a frequent speaker in recent years on topics related to the future of cataloging, particularly with regard to technical services workstations and the famous "More, Better, Faster, Cheaper." He organized and led a series of eight coast-to-coast institutes, sponsored by ALCTS, LITA, LAMA, and ACRL on "Technical Services Workstations: the State of the Art of Cataloging." Dr. Kaplan edited the 1997 monograph *Planning and Implementing Technical Services Workstations*, which was published by ALA Editions, and he contributed to the 1996 ARL SPEC *Kit on Technical Services Workstations* (TSW).

Cataloging Directorate Home Page

Library of Congress Home Page

In 1997, Dr. Kaplan received the ALCTS Best of LRTS award for his article *"Technical Services Workstations: A Review of the State of the Art."* and in 1998, he received the LITA/Library Hi Tech Award for his body of work over the previous five years that showed "outstanding achievement in communicating to educate practitioners within the library field in library and information technology."

Full text of paper is available

Summary:

In two centuries we have come a long way in the construction of our bibliographic catalogs: from book to card catalog to comfiche to OPACs, and now the Web. The catalog data that underpinned those presentation devices--what some people refer to as real metadata as opposed to naive metadata created by non-catalogers or designed to describe newer types of electronic materials--has changed as well, but seemingly with fewer phases. If we ignore the issue of particular metadata standards and keep to AACR in its various iterations, and now the Dublin Core-inspired metadata standards, then what I am concerned with bluntly is with the marriage of metadata standards and presentation. In leading this panel, I first envisioned a group of vendors talking about the catalog record as a dynamic entity and their role in creating it. I was originally intrigued by several varieties of cataloging or cataloging enhancements that are becoming more and more significant to us and our patrons. Three views, briefly:

Aggregators and aggregations: Like their printed or microform counterparts of the 1970s and 1980s, aggregations, principally of serials, threaten to overwhelm us. Decisions are made to purchase large electronic sets and then we in technical services are left holding the virtual bag trying to offer access.

Ancillary data: The oldest example, probably, is the table of contents pioneered by Blackwell North America and others in the late 1980s and early 1990s. As envisioned, a library would subscribe to some or all of a set or purchase TOC data for individual titles. Now, TOC data is but one piece of the constellation in a galaxy of similar constellations: why not add back of book indexes, author portraits, summaries, or book reviews?

Then there is the advent of metadata for electronic books, the key to discovering and ordering from vendors such as netLibrary. This data may reside in our catalogs, but even more appropriately on the Internet, and appeal not only to librarians, but also even more directly to the end user. However, the real opportunity that comes from all this is a relatively new adjunct to the Internet and the Web. This is the technology created by Herbert van de Sompel and his colleagues at the University of Ghent. Called SFX for Special Effects (not to be confused with SFX technology for delivering audiovisual resources over the Web), it is nothing short of a revolution in how we should envision research on the Web. SFX is a framework for context-sensitive linking between Web resources. It is the means to uniting or linking disparate, heterogeneous electronic resources such as abstracts and full text, all the while keeping in mind the context in which the user works and that some sources of data may be institutionally more appropriate for that user than others. It also has the ability to link to related subjects. This is truly exciting, yet I am struck by the notion that we have hit on one of the holy grails of research. The Holy Grail is that of "seamless interconnectivity." To back up a step, this technology is seamless only because the metadata exists as seams in an information architecture. SFX then takes the seams one step further and turns them into a library-defined seamless whole (WHOLE, not HOLE).

Amira Aaron,

commentator

Director of Marketing and Programs
Faxon, RoweCom's Academic and Medical Services
15 Southwest Park
Westwood, MA 02090

About the commentator:

In her role as Director of Marketing and Programs at Faxon/RoweCom, Aaron serves as a primary liaison to the academic library marketplace. She participates in strategic planning and product development for the academic and medical communities and heads the Academic Client Advisory Board.

Aaron came to Faxon from Blackwell's Information Services Group, where she was the Electronic Services Product Manager for serials management and push technology products. Prior to Blackwell's, she served as the Coordinator of Library Automation and Product Development at Readmore, specializing in the development of bibliographic interfaces and Internet services. Aaron also has significant experience in academic libraries, having held several key technical services and automation positions at the UCLA library system, including Head of Continuations Cataloging and Associate Head of Technical Services. She also served as Associate Director for Systems and Planning at the Massachusetts Institute of Technology libraries.

Aaron holds an MLS from UCLA and completed coursework in the university's Ph.D. program as well. She served as the co-chair of the SISAC Education and Publicity Subcommittee for several years. Aaron is currently chair of the ALA ALCTS Serials Section Research Libraries Discussion Group and is a member-at-large of the ALCTS Serials Section Executive Committee. She has organized several noteworthy professional conferences and is a frequent speaker on library technology and serials management topics.

Full text of commentary is available

Jeff Calcagno,

commentator

Director of Sales and Customer Support
Syndetic Solutions, Inc.
7521 S.W. Garden Home Rd.
Portland, OR 97223

About the commentator:

Jeff Calcagno directs Syndetic Solutions Library Sales and Customer Support group, which provides an expanding array of catalog enrichment services to libraries. He has an MLS from the University of Washington and has been a well-known figure in the library technical services industry since 1986. Prior to the formation of Syndetic Solutions, Jeff was a Senior Account Manager in the Technical Services Division at Blackwell's where he provided counseling and project management support for many large research universities, public libraries, consortiums, and libraries and national bibliographic networks within the Pacific Rim. In addition to library sales and support, Jeff manages Syndetics' consulting services, providing technical and marketing support for library networks and commercial bibliographic service providers.

Full text of commentary is available

Summary: Library catalog users are often web users. They have experienced, and continue to utilize, enhanced bibliographic information from the web which gives rise to heightened expectations for the library catalog. This paper outlines some of those perceived expectations and provides information on what types of enrichment data libraries should plan to receive from vendors. The paper also reviews several attributes of the data, in what manner libraries presently receive the data, and concludes by noting several implementation issues which must be addressed by libraries and vendors of enrichment data.

Dr. Lynn Silipigni Connaway, commentator

Vice President of Research & Library Systems
netLibrary, Inc.
3080 Center Green Drive
Boulder, CO 80301

About the commentator:

As Vice President of Research and Library Systems, Lynn Silipigni Connaway is responsible for directing internal research and plays a critical role in the creation of the company's information search interface for the library community. She also oversees the collection development and cataloging teams.

As a former professor of library and information science, Dr. Connaway's area of expertise is in the field of information cataloging and classification. Prior to joining netLibrary, Dr. Connaway served as Director of the Library and Information Services Department at the University of Denver, where she taught several courses in library and information science. During her tenure, she conducted research on the subjects of organization and access of electronic documents, as well as the education of information professionals.

Dr. Connaway has served on the faculty of the University of Missouri, Columbia and as a lecturer at the University of Wisconsin, Madison, and is a frequent speaker at national professional meetings and conferences.

Dr. Connaway received a Ph.D. in Library and Information Studies from the University of Wisconsin, Madison, a Master of Library Science degree from the University of Arizona, and a Bachelor of Science degree in Library Science from Edinboro State University. She has been actively involved in numerous committees within the American Library Association, the American Society for Information Science, and the Association of Library and Information Science Educators.

Full text of commentary is available

Summary:

The emergence of electronic books (eBooks) in libraries has brought new opportunities and new challenges. The opportunity to provide access to full-text eBooks brings the challenges of making them available through standard library practices and systems. The integration of eBooks into

libraries' collection development and acquisitions processes and into online public access catalogs (OPACs) requires the cataloging of these materials. Some of the challenges identified are adhering to cataloging standards; integrating other industry standards and schemes; establishing and updating bibliographic links; classifying; reporting statistics; adapting work flow processes; and training staff and patrons.

As an eBook provider, netLibrary, Inc. has been involved in the selection, creation, cataloging, and distribution of eBooks. This involvement has given netLibrary staff a first-hand look at some of the challenges associated with eBook access, however, challenges often bring new opportunities. This is the ideal time for librarians, publishers, technology providers, eBook providers, and bibliographic utitilies to work together to develop standards and processes that will meet each group's needs, but most importantly, will meet the needs of the individuals who use electronic resources.

Library of Congress
December 21, 2000
Comments: lcweb@loc.gov

Exploring Partnerships:
What Can Producers and Vendors Provide?

Michael Kaplan
Director of Product Management
Ex Libris (USA), Inc.

Final version

In two centuries we have come a long way in the construction of our bibliographic catalogs. We have evolved from the book style catalogs of a former century that were also made famous in this century by the enormous catalogs such as those of the British Library and our own National Union Catalog-surely the book catalog/union list to end all catalogs-to card catalogs that comprised 1000s of drawers at large institutions to microfilm or microfiche and thence to comfiche and then to online OPACs. While I can say that all in a single list of nouns with a grammatically syndetic structure (that is to say, an 'and-ed' string), the fact is that this small list-book, card, film, fiche, computer output microform, and online-represents a credible chronological period. Call it 150 to 200 years, give or take. Another way to measure it is to say that is encompasses about 30% of the time that has elapsed since the invention of the printing press with moveable type.

And this only addresses the issue of the form or format of the individual catalog records and the catalog itself. I'd like you to view the catalog records as analyzed pieces of data, discrete but interconnected, and the catalog proper as the container for the actual catalog records. The individual records linked together in what we used regularly to call a 'syndetic' structure are what make a catalog out of discrete bits of data.

On this level we have created the actual standards and protocols of the catalog records that really are the infrastructure of the catalog itself. These are the standards upon which individual records interconnect and hold together as a unified, coherent whole. These codes, both the historical codes and the present-day codes, are the standards and protocols of the catalog just as DNA is the protocol of the genes that make up all living beings. Some among us who are a bit jaundiced about our ability to adapt might claim that the codes are more immutable by far than DNA-and even sometimes as hard to understand. Change happens slowly-but, unlike with life forms, never by accident. There are no accidental mutations in our world. Our catalog code mutations are deliberate and are subject to national and international standards bodies that give them their seal of approval only after due consideration.

It is worth observing that the catalog formats and the catalog codes with which we are most familiar were developed in an overlapping fashion. Most of us would probably consider the late 19th and the 20th centuries as the period that comprises the age of substantial catalog form development and the time when bibliographic control or, at any rate, the age when meaningful bibliographic cataloging codes, developed.

But neither time nor technology stands still and now we have the World Wide Web. The Web. No one needs to ask what you mean; there is only one Web and it is 'the' Web. You don't even need to capitalize the 'T' in 'the' nor, I think, pronounce 'the' as 'the' with a long 'e'. That is what the Channel 56 television news does in Boston to show that they are the first and original 10 PM newscast: they call themselves "The Ten O'clock News." The Web is surely different, however. It is not Charlotte's Web, nor anyone else's Web. It belongs to all of us, if it can be said to belong to anyone. After all, the Internet is famous for its loose governance structure. That is what makes it so exciting and so revolutionary, but it is also what has come to make life so challenging, not to say difficult, for those of us whose lives are centered on the concept of bibliographic control.

I hope that those of you may have had training in classics, as Eric Jul and I did, or at least in mythology, will agree with me when I say that I find parallels between the sudden efflorescence of the Web and the birth of the goddess Athena. According to Hesiod, Athena was born fully-grown from the head of Zeus. Pindar expounded on that a bit further and tells us that Hephaistos struck open Zeus' head with an axe to release Athena. The Web is not yet a decade old and its precursors go back considerably further. Yet it has come upon us so suddenly and with such overwhelming force that it can be thought to have sprung almost as fully grown and fully armed as Athena did at her first appearance. For a technology that young it has attained a tremendous level of maturity and acceptance, but bibliographic standards as we know them are not yet among its major accomplishments. That being said, I feel safe in saying it has certain characteristics of a young and rebellious child in that it refuses to be governed by the long-standing rules of its elders. Unfortunately for those of us whose lives revolve around issues of bibliographic control, its growing pains are our parental pains.

In the early 1990s I was a member of OCLC's Cataloging and Database Services Advisory Committee when OCLC undertook an experiment, the InterCat project, to catalog electronic resources. Electronic lists and journals were still few enough in number that it seemed possible to actually control this new realm in a semi-traditional fashion. It was at this time that the MARC field 856 was defined for the URL. Not too many years later the first of the Dublin Core conferences was held, and by then it was abundantly clear that we had entered into a new dimension where the traditional no longer held sway. In talking about cooperative relationships in building metadata I need to note the absolutely pivotal position of OCLC and its Office of Research in this, not only because of this InterCat experiment and the development of the 856, but also because from the beginning this experiment was conceived of as cooperative in nature. Martin Dillon and Eric Jul were visionary within OCLC's Office of Research then, and they continue to help guide us today. In a hallway conversation at this summer's ALA conference Martin and I were talking about the program entitled "Is MARC dead?" We were talking about XML and Dublin Core, and he observed (perhaps quoting Fred Kilgour?) that the struggle to define and refine and redefine Dublin Core has taken longer than World War II did to fight and win!

On the face of it that is true and you have got to wonder about our sense of priorities and perspective. Perhaps a rephrasing the famous Latin phrase Vita brevis, ars longa (Life is short, but Art is long) is called for: Data brevia, regula longa, or, It only takes a moment to create a record, but the route to standards is long! Finally, I would alter Martin's hallway comment about the Dublin Core and compare it not to World War II but to the Hundred Years' War. Like the Hundred Years' War, MARC and AACR2 and Dublin Core have their adherents and their political and religious believers. Like many of the struggles in bibliographic control, as that over Full- vs. Minimal- vs. Core-level cataloging, the struggle resolves itself to a common denominator of ideology.

Let me turn back to the subject at hand. The data that underpins our catalogs has a long and distinguished history. It is what some people would refer to as real cataloging or real metadata. Of course, very few of us had ever heard, let alone used, the term metadata until cataloging of Internet resources became a part of our job descriptions. I suggest that the new term was adopted partly because new standards became an issue and so a new term seemed appropriate, but also because the issue of providing descriptive and other sorts of data about them had been co-opted by computer scientists. In her keynote presentation at the ALCTS Conference "Metadata: Libraries and the Web" at last summer's ALA Annual Conference, Jennifer Younger reported on a brief survey of online literature for use of the term metadata. Searching the Web of Science she found some 279 hits on the term, of which the oldest was from 1982. I myself had previously tried searching for it in the online Oxford English Dictionary and Merriam-Webster, but to no avail. It has not yet entered popular parlance to the degree that it is recognized by those authorities.

So, if scientists first made use of the term to refer to categorization and classification and then passed it on to the library science community, I would not be surprised. We have had and continue to have our challenges in dealing with a scientific community that is not grounded in our discipline and that is amazed that we had a tradition of authority control long before they ever found a need for it. I suspect they concocted the term 'metadata' because they did not know that we already had a perfectly good name for what they were engaged in doing. By that I mean 'cataloging', of course.

By now we have all absorbed the concept of metadata. I am not sure that we have all come to recognize it as a semi-Aristotelian concept, however. Aristotle wrote the Physica and then the Metaphysica. For him it was an issue of works on physics and the things that came after that. In Greek it was really the Physica and the Meta ta Physica, the Physics and the AfterPhysics. We might as well speak of CatalogData and MetaData or TraditionalData and NaïveData or Metadata. I cannot take credit for NaïveData; it is what my good friend and colleague from Yale, Matthew Beacom, likes to call the New MetaData. By Naïve Metadata he is referring to cataloging created by non-catalogers and designed to describe newer types of electronic materials.

It is another perspective on what Elizabeth Mangen of the Geography and Map Division here at the Library of Congress said at the ALCTS metadata conference: "Metadata is a complement to cataloging data, not a replacement for it."

As comforting as it would be to espouse such sentiments, I am afraid that I cannot. If Cicero were here with us today, he might well exclaim, "O tempora, o mores"-"Oh, the times, oh, the customs." I cannot. It is simply too atavistic to pretend that the old solutions are still viable and that the world is not changing in revolutionary ways around us. Nor that we can hunker down in our bibliographic bunkers and go it alone. We cannot.

Be that as it may, I suspect that if we as a profession had been more nimble on our feet, if we had been quicker to change and to recognize the profound changes that were about to confront us, and if we could have known with the power of hindsight the absolute explosion of data that was to engulf us, we might have tried to accelerate our standards processes to encompass these new data types. We may have, but I am not convinced that we could do it. Standards-setting bodies by their nature are not prone to moving quickly. I remember one of the very early formative meetings of the Cooperative Cataloging Council even before it became the Program for Cooperative Cataloging. A group of forward-thinking individuals was concerned with adapting our cataloging codes more quickly to the challenges of Internet resources and seeing us be proactive, not reactive by a factor of years.

We have come a long way, but we still too much resemble, to my mind, the legal profession. It is obvious that the legal profession lags far behind technology in societal and ethical issues, whether it is the length of time it has taken to establish the validity of digital signatures or the plethora of medical conundrums that advances in medical science force upon us. While I recognize the necessity of affirming and reaffirming our roles as custodians and catalogers in an historical and intellectual framework, we will lose our role unless we learn to swim faster than ever before. It is either that or we drown. Or, as I suggest, we learn to engage in cooperative swimming. We must, because we are no longer swimming in a small, well-defined, and clearly bounded swimming pool. We are attempting to swim in an ocean, and one being constantly replenished with enormous rivers of data every second of every day.

I am reminded of a NELINET annual meeting a few years ago. Marshall Keys, the then Executive Director, was giving his annual state of NELINET address. He told us of a job description for an institution in North Carolina, I think it was, that included this statement: "Successful candidate does not have to be able to walk on water, but definitely must be a strong swimmer." I think that is an apt description of what we are trying to accomplish in technical services today.

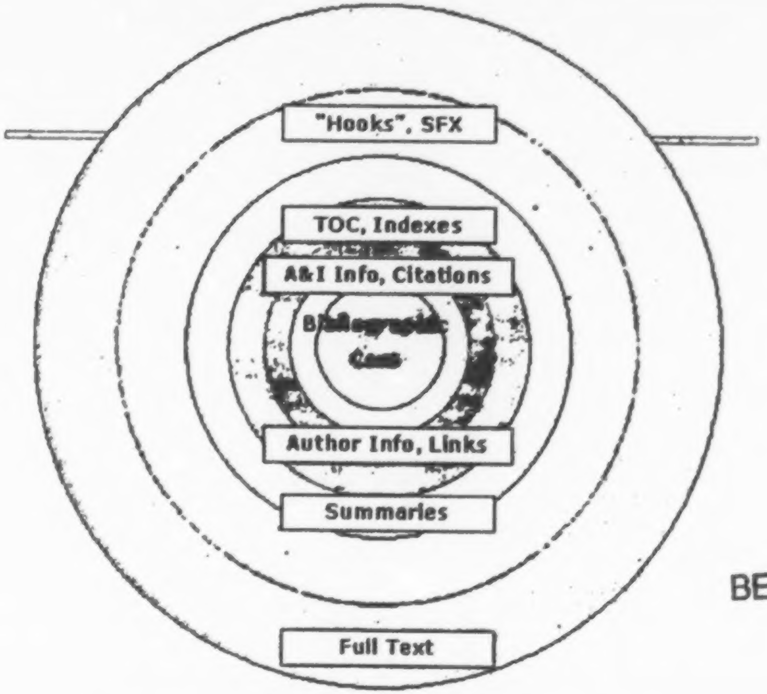
Now, if I am candid and truthful, I will admit to you that in the academic library circles where I have spent most of my professional life we were already in danger of drowning a long time ago. It is not the Internet phenomenon that alone threatens to overwhelm us. I spent many years at a large, well-known institution in Cambridge, Massachusetts, at the opposite end of Massachusetts Avenue from MIT-the very place that Priscilla Caplan and Robin Wendler and I met-that never managed to process, let alone seriously catalog, anything like the major part of its enormous annual intake of new materials. As long ago as 1992 I was venting on these subjects in a talk entitled "Minimal-level Cataloging and Other Solutions to the Backlog" in a seminar on "Cataloging in the '90's," sponsored by New England Technical Services Librarians and NELINET. I may be guilty of revisionist history, but I don't seem to recall advocating minimal-level cataloging at that time, but I did strongly recommend a more fundamental reassessment of what we meant by copy cataloging and inter-institutional cooperation. These became, in fact, some of the basic tenets of the Program for Cooperative Cataloging. Minimal-level cataloging concepts, on the other hand, led by various pathways to the definition of the Dublin Core. To revisit a Reagan-era mantra, I believe that I advocated the concept of 'trust but verify' in the context where verification was predicated on machine-assistance from both the client and the server levels.

So, if we were already in danger of drowning in a sea of print-based materials by 1992, how can I characterize our state in 2000? Really drowning? Really, really drowning? Or really, really in need of help and new helpers? How much drowning in a sea of uncataloged or uncontrolled materials can any one individual or institution endure before they are dead, or really, really dead?

Let me turn back to the issue of standards, new standards, and the information landscape that stretches before us.

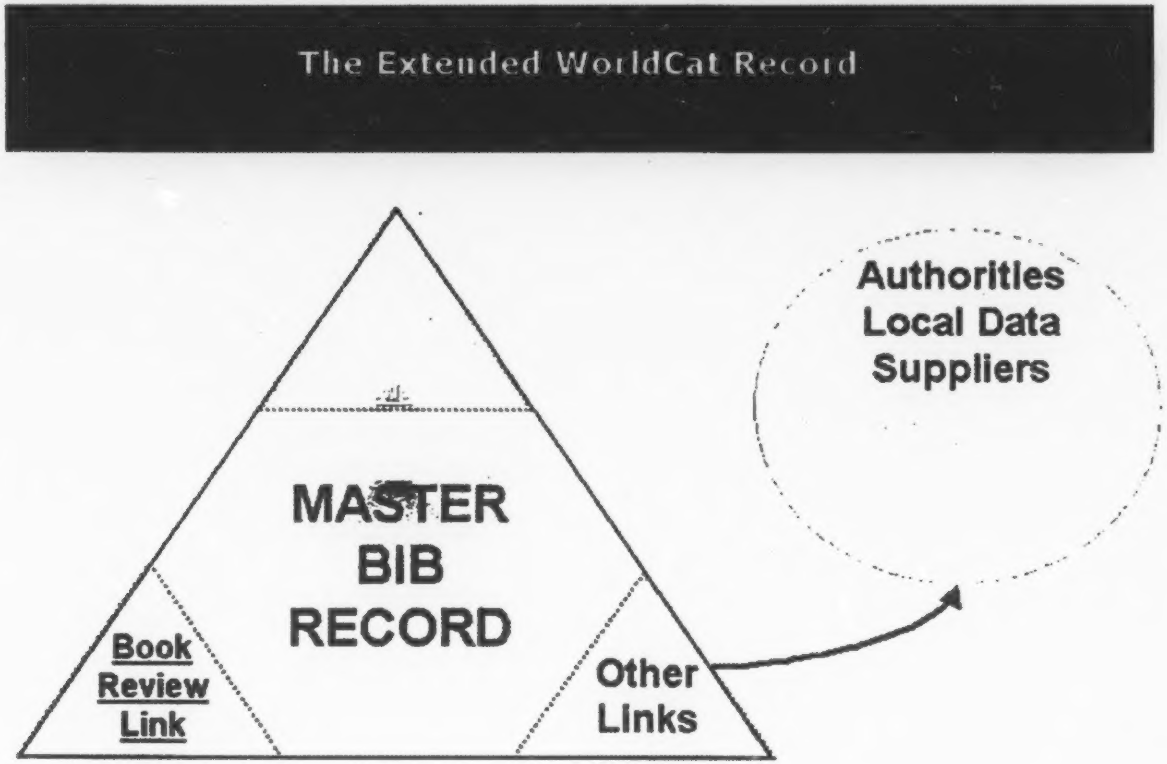
In the beginning of this story we had the ALA data standard, then the Anglo-American standards, AACR, known in retrospect as AACR1, then AACR2, finally AACR2 with its various revisions. The Germans have their RAK. On the other side are MARC21 and other data structures: the Dublin Core metadata standard is the most prominent these days. But there is an entire host of other standards out there as well, whether it is the Visual Resources Association (VRA) Core, the Enhanced Archival Description (EAD), the Text Encoding Initiative (TEI), the Computerized Interchange of Museum Information (CIMI), or the Records Export for Art and Cultural Heritage (REACH). There are geospatial standards and there non-Anglo-American standards-the German MAB and MAB2. We Anglophones need to accept the fact that, while English is at the center of the modern information network, it is not universal, and we need to reach out and be linguistically inclusive and find ways to map and marry different national, linguistic data standards to our own every bit as much as we need to consider different data structures. As hard as it has proved to harmonize the former USMARC and CanMARC into MARC21, and with UKMARC still to join the other two, reaching across linguistic boundaries is even harder. I am vitally intrigued and hopeful that someday we will see a dynamic structure that allows all linguistic communities to participate in metadata creation, both bibliographic and authority, and that we can have a system where we Anglophones can use a heading such as Cologne (Germany) and the Germans can use Köln (Deutschland) and intelligent systems can substitute one for the other based on individual and institutional contexts. In essence this is Barbara Tillett's thesis. But I am not naïve about the challenges-I do not expect to see such a system anytime soon, perhaps not even this decade.

How well do you remember the old style book catalog? It was very one-dimensional and extraordinarily static. The National Union Catalog and its regular cumulations were like that. There tended to be but a single access point, and that was it. With card catalogs we developed the ability to photocopy multiple cards, create additional access points, and maintain a multiplicity of complementary files. A declining number of us-a number becoming fewer every year-enjoyed the rigors of indentions, spacing, and typing in red! As I like to recall, one of the inhibitions on multiple access points was the need for the photocopying and manual typing. OCLC and RLIN cards sets put that particular inhibition to rest. In fact, I suggest that catalog records endured a gradual increase in length over time because the computer-assisted nature of card production made that aspect of catalog generation easier and easier. However, fewer and fewer of us have recollections of dealing with multiple catalogs (union, author, title, subject, dictionary). Those cards that have all but disappeared-and, as I like to say, to a humiliating end! since they have lost their lot in life as scrap cards to the ubiquitous yellow stickies. But let us not forget that those cards offered the first meaningful ability to cross-link and trace from one card to another, and from one alphabetical arrangement in the catalog to another.



BEST COPY AVAILABLE

What confronts us then is finding a way to reconcile-I will not say to 'unite', but to 'reconcile'-different metadata structures and standards and to develop sensible presentations of them. I will try to keep this simple and presume that the metadata structures are Anglophone-centric, not in their content but rather in their bibliographic and authority structures. Now let me take it one step further and say that the future record, whether you prefer to call it a catalog record or a metadata record, will not be one-dimensional or static. Rather, it will be multi-faceted and dynamic. It will be composed on-the-fly of a variety of different "metadata-lets": the traditional bibliographic description at its core, but with a number of concentric circles associated with it and including such information as citations, reviews, dust jacket illustrations, author information, links for delivery or ordering, etc. I think it fair to say that we have been moving gradually over a long period of time toward that vision, but technology is only now getting us to a point where it can be realized. Widespread use of A&I and citation databases and the creation of table of contents databases separate from the MARC record were the first glimmers of enhanced records. Alive to the possibilities, OCLC has now presented sketches of its planned successor service to WorldCat that will be based on a CORC standard and tentatively called 'eXtended WorldCat':



The Extended WorldCat Record

Now, when I was asked to lead this panel by John Byrum more than a year ago, I was Associate Dean of Libraries and Director of Technical Services at the Indiana University Libraries-Bloomington. Since that time, as you know, I have left a lifetime in academia and taken the position of Director of Product Management for Ex Libris (USA), Inc. As I envisioned the panel we would have a group of vendors talking about the catalog record as a dynamic entity and their role in creating it. What do I mean by a dynamic record? It is a record that has a core consisting of a traditional bibliographic description, but it is regularly enhanced and/or refreshed by a series of elements that we have not routinely

BEST COPY AVAILABLE

associated with the record. These elements will be provided to us as part of a regular bibliographic commerce conducted on the fly. In addition, we will benefit from regular infusions of data files from other sources, data that will accompany regular infusions of digital text files.

I want to make clear that I am not talking about outsourcing our bibliographic responsiblities in a wholesale sense, but there is an element of it here and I want to talk about it in a non-traditional sense. I want to discuss it as involving either discrete parts of the record, on the one hand, or of aggregations, on the other. Where parts are at issue, I want to talk of outsourcing the way General Motors would when it buys transmissions from one supplier and engines from another, or better, when it buys optional, value-added parts of the automobile-roof racks or moon roofs-from outside vendors. Engines and transmissions are essential to automobiles just as a core bibliographic description is essential to the bibliographic record. I want to talk about purchasing parts of records from diverse sources that are in the best position to manufacture them, provided of course they assure us of the quality control that we have every right to demand.

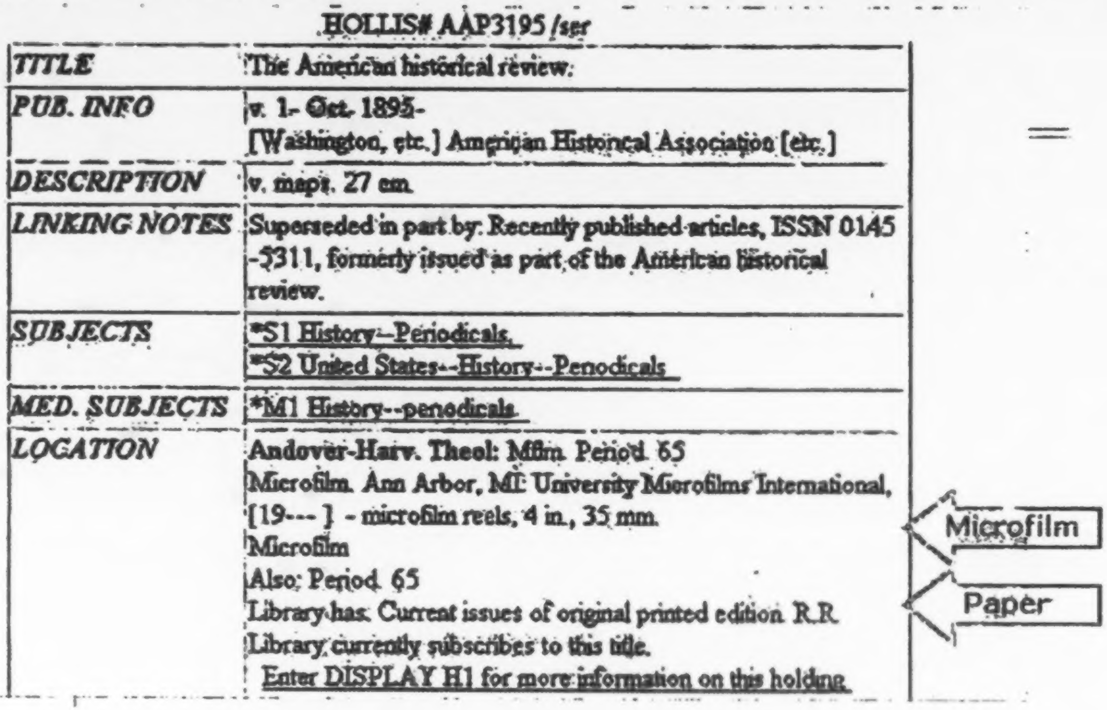
I was originally intrigued by several varieties of cataloging or cataloging enhancements that are becoming more and more significant to us and to our patrons. Let me offer to you four views of them.

- The first concerns aggregators and aggregations.

Like their printed or microformat counterparts of the 1970s and 1980s, electronic aggregations-principally of serials-threaten to overwhelm us. In their size and complexity they rival the major microform sets that we purchased during the past two to three decades. To some degree we managed bibliographic control over those sets by cooperative cataloging and bulk purchases and batch loading. In their absence, we usually made do with a printed index to the set. Unfortunately, few libraries are left that have the time or the staff to handle catalolging of these massive sets any longer. It is my perception that a collection-level record in the catalog for databases such as JStor or Project Muse or Ovid or any of the others have little utility beyond providing a place for recording payments. While some users might approach their journal searching on this level, most users clearly are after specific titles and not the database in which they are contained. (More on this below.) Into this void the publishers-for example, Bell and Howell and Primary Source Media-have now, thankfully, stepped and are offering to sell us the cataloging copy that goes with the sets. I would frankly be happier if they recognized that electronic copy for these sets is essential and the sets should be priced with the copy included and above all did not treat the cataloging copy as an add-on available at additional cost.

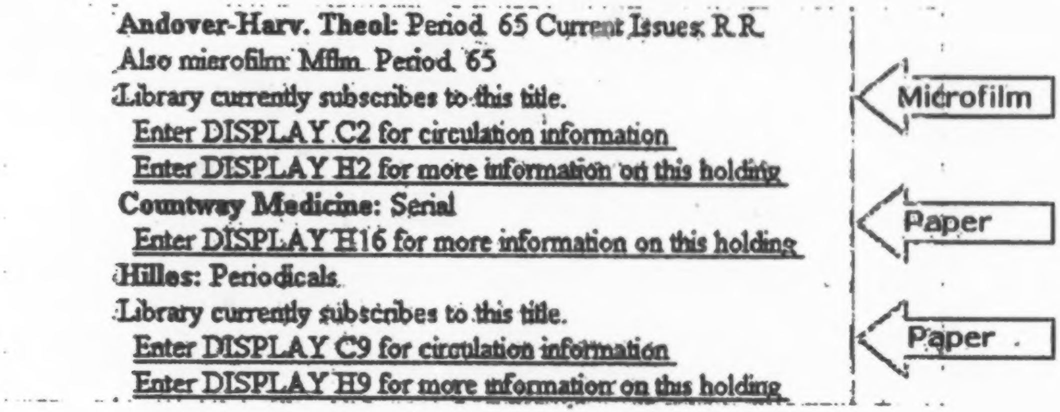
Where electronic, primarily serial aggregations are concerned, decisions are routinely made to purchase large electronic sets and then we in technical services are left holding the virtual bag trying to offer access. The Program for Cooperative Cataloging is setting basic standards and enticing the aggregators to provide us with copy, and I applaud EBSCO and others who are participating in our experiment. But, when the individual titles come and go at the speed of light, even their aggregator creators may not always know what is in the aggregations. In fact, I can tell you, without naming names, that one of the largest aggregators welcomed the PCC initiative precisely because it gave them an excuse to try to inventory their offerings. It is essential for us to make the issue of regular, low-cost or no-cost access to electronic catalog records a matter of competitive market advantage in our dealings with the vendors of aggregated content.

It should by no means be beyond the ability of aggregators to provide us with this data, and the best incentive for them is likely an economic one if libraries, principally collection development officers, are willing to wield it as a trump card. I have in mind that electronic records for all sets be provided in standard form by publishers as a part of their set and that their presence be considered an essential part of the evaluative process that precedes buying them. I have invited Amira Aaron, Director of Marketing and Programs for Faxon/RoweCom, to help us think through all these challenges.



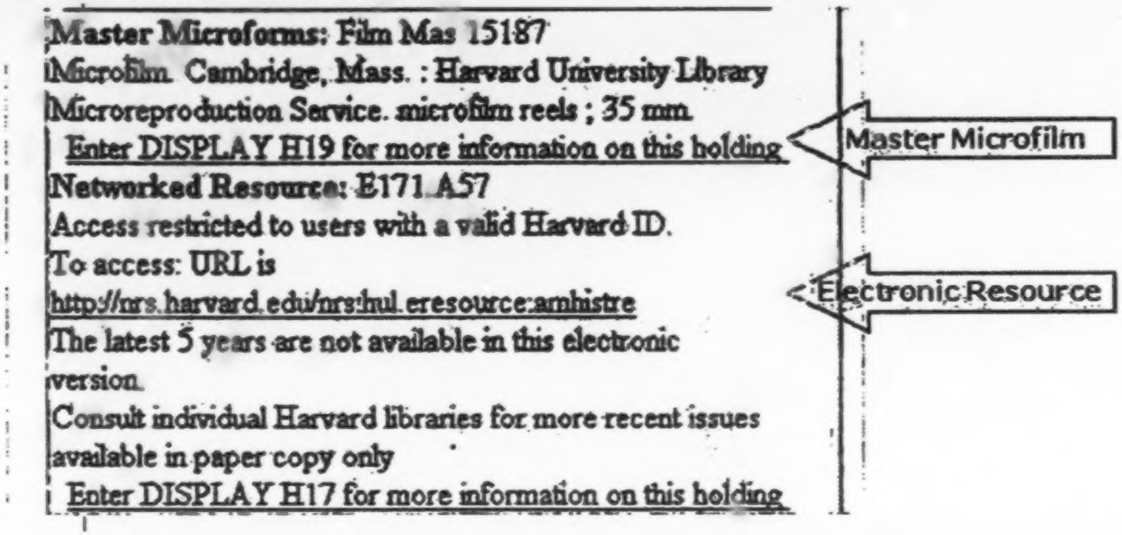
HOLLIS# AAP3195 /ser

TITLE	The American historical review.
PUB. INFO	v. 1- Oct. 1895- [Washington, etc.] American Historical Association [etc.]
DESCRIPTION	v. maps. 27 cm.
LINKING NOTES	Superseded in part by: Recently published articles, ISSN 0145-5311, formerly issued as part of the American historical review.
SUBJECTS	*S1 History--Periodicals. *S2 United States--History--Periodicals
MED. SUBJECTS	*M1 History--periodicals.
LOCATION	Andover-Harv. Theol: Mflm. Period. 65 Microfilm. Ann Arbor, MI: University Microfilms International, [19---] - microfilm reels, 4 in., 35 mm. Microfilm Also: Period. 65 Library has: Current issues of original printed edition. R.R. Library currently subscribes to this title. Enter DISPLAY H1 for more information on this holding



Andover-Harv. Theol: Period. 65 Current Issues: R.R.
Also microfilm: Mflm. Period. 65
Library currently subscribes to this title.
Enter DISPLAY C2 for circulation information
Enter DISPLAY H2 for more information on this holding
Countway Medicine: Serial
Enter DISPLAY H16 for more information on this holding
Hilles: Periodicals
Library currently subscribes to this title.
Enter DISPLAY C9 for circulation information
Enter DISPLAY H9 for more information on this holding

BEST COPY AVAILABLE

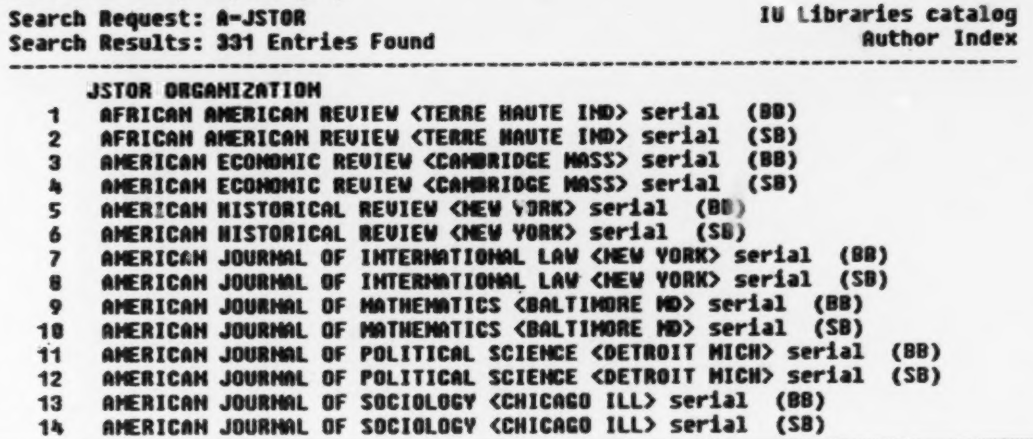


Master Microforms: Film Mas 15187
Microfilm. Cambridge, Mass. : Harvard University Library
Microreproduction Service. microfilm reels ; 35 mm.
Enter DISPLAY H19 for more information on this holding
Networked Resource: E171.A57
Access restricted to users with a valid Harvard ID.
To access: URL is
http://nrs.harvard.edu/nrs:hul.eresource:amhistre
The latest 5 years are not available in this electronic version.
Consult individual Harvard libraries for more recent issues available in paper copy only
Enter DISPLAY H17 for more information on this holding

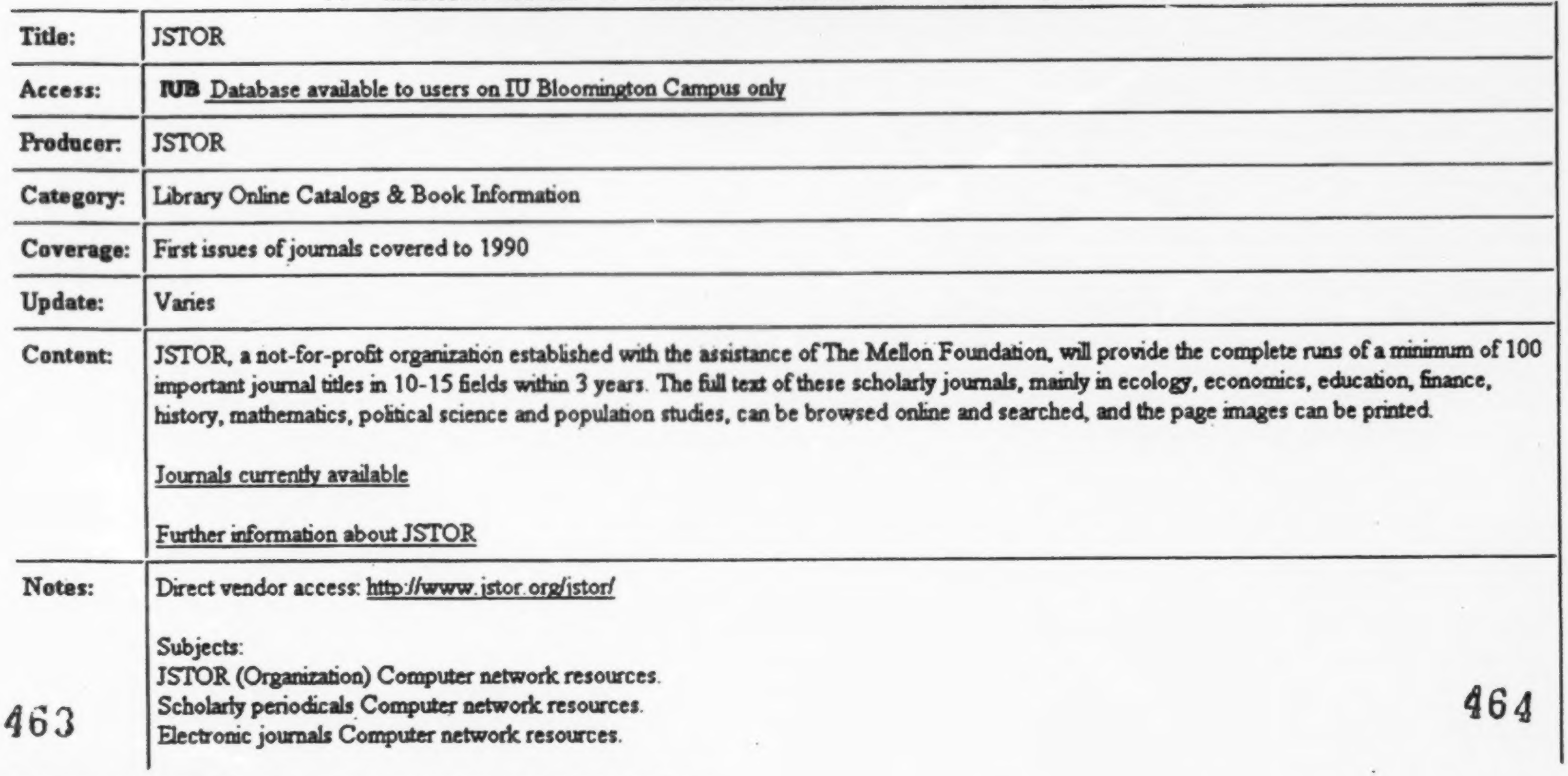
Having had experience with purchasing some hundreds of thousands of major microform-style records while Head of Database Management at Harvard University in the early 1990s and then as Director of Technical Services at Indiana University-Bloomington in the late 1990s, I can tell you that there are substantial standards and data-loading issues to overcome. One of the principal challenges we face with both major microform sets and electronic aggregations is that of multiple versions and the single-record approach to the catalog. The challenge, as I see it, is more static in the microform arena than that of the electronic aggregation, and I personally favor the Mulver approach long held at Harvard University. Yet, while it can be relatively easy to match and load the microform copy to the record for the paper original for monographs and then to forget about it, we have no such luxury where the paper and digital worlds collide. That is because a very high percentage of the digital content in the typical aggregation comes and goes with alarming frequency. I have a very serious challenge to pose to the aggregators and to the system vendors among us, and it is this:

Libraries have a pressing need to develop a cradle-to-grave approach to handling electronic sets. That means obtaining the corresponding electronic records from the aggregator and receiving regular maintenance updates to the sets. Maintenance would ideally include additions, deletions, changes in coverage, etc. Now, as fond as I am in a theoretical sense of the Mulver or one-record approach also to electronic journals, I frankly do not see any simple way for us to allow the aggregator's data file to reach deep down into our integrated library systems and to touch data on this level if the data is buried within a Mulver-style record. It is simply fraught with too many difficulties. So I reluctantly conclude that, if we want to look to a hands-off, computer-to-computer data interchange, we need to keep the basic building blocks of that exchange as simple as possible. From the perspective of a former Director of Technical Services in one of the so-called Big Heads libraries, I have to admit that managing large serial aggregations is an impossibility at the local level without firm and decisive actions from the aggregators and help from the ILS vendors. I am desperate for an EDI-type solution for all aspects of aggregations that drills right down into the local catalog to solve this problem that I see as growing increasingly intractable.

So let me add a specific recommendation to this complaint. Contrary to what I said earlier about preferring the Mulver solution to disparate formats, I have reluctantly concluded that a partnered solution here will mean keeping the electronic journal records separate from their paper counterparts. That way an incoming record, particularly an incoming maintenance-level record, can be programmed to behave in predictable ways vis-à-vis the existing database record. This puts my public services preference at odds with my technical services solution, but I believe the solution that best serves technical services and that provides the best, easiest, and most-up-to-date access will ultimately prove the best public services solution. Besides, it might prove possible to then merge records for the public view that are kept separate in the technical services components of our catalogs.



```
Search Request: A-JSTOR                              IU Libraries catalog
Search Results: 331 Entries Found                            Author Index
--------------------------------------------------------------------------
   JSTOR ORGANIZATION
 1  AFRICAN AMERICAN REVIEW <TERRE HAUTE IND> serial  (BB)
 2  AFRICAN AMERICAN REVIEW <TERRE HAUTE IND> serial  (SB)
 3  AMERICAN ECONOMIC REVIEW <CAMBRIDGE MASS> serial  (BB)
 4  AMERICAN ECONOMIC REVIEW <CAMBRIDGE MASS> serial  (SB)
 5  AMERICAN HISTORICAL REVIEW <NEW YORK> serial  (BB)
 6  AMERICAN HISTORICAL REVIEW <NEW YORK> serial  (SB)
 7  AMERICAN JOURNAL OF INTERNATIONAL LAW <NEW YORK> serial  (BB)
 8  AMERICAN JOURNAL OF INTERNATIONAL LAW <NEW YORK> serial  (SB)
 9  AMERICAN JOURNAL OF MATHEMATICS <BALTIMORE MD> serial  (BB)
10  AMERICAN JOURNAL OF MATHEMATICS <BALTIMORE MD> serial  (SB)
11  AMERICAN JOURNAL OF POLITICAL SCIENCE <DETROIT MICH> serial  (BB)
12  AMERICAN JOURNAL OF POLITICAL SCIENCE <DETROIT MICH> serial  (SB)
13  AMERICAN JOURNAL OF SOCIOLOGY <CHICAGO ILL> serial  (BB)
14  AMERICAN JOURNAL OF SOCIOLOGY <CHICAGO ILL> serial  (SB)
```



Title:	JSTOR
Access:	**IUB** Database available to users on IU Bloomington Campus only
Producer:	JSTOR
Category:	Library Online Catalogs & Book Information
Coverage:	First issues of journals covered to 1990
Update:	Varies
Content:	JSTOR, a not-for-profit organization established with the assistance of The Mellon Foundation, will provide the complete runs of a minimum of 100 important journal titles in 10-15 fields within 3 years. The full text of these scholarly journals, mainly in ecology, economics, education, finance, history, mathematics, political science and population studies, can be browsed online and searched, and the page images can be printed. Journals currently available Further information about JSTOR
Notes:	Direct vendor access: http://www.jstor.org/jstor/ Subjects: JSTOR (Organization) Computer network resources. Scholarly periodicals Computer network resources. Electronic journals Computer network resources.

463

464

JOURNALS CURRENTLY AVAILABLE

View currently available journals, alphabetically or by subject.

Moving Wall information

African-American Studies | Anthropology | Asian Studies
Ecology | Economics | Education | Finance
History | Literature | Mathematics | Philosophy
Political Science | Population Studies | Sociology | Statistics

General Science Collection

Institutions participating in the General Science Collection
are noted on our Participants page.

African-American Studies

African American Review Vols. 1-30, 1967-1996
Moving Wall: 3 years
Journal URL: http://www.jstor.org/journals/10624783.html
Publisher: Indiana State University
(continues Black American Literature Forum)
Journal URL: http://www.jstor.org/journals/01486179.html

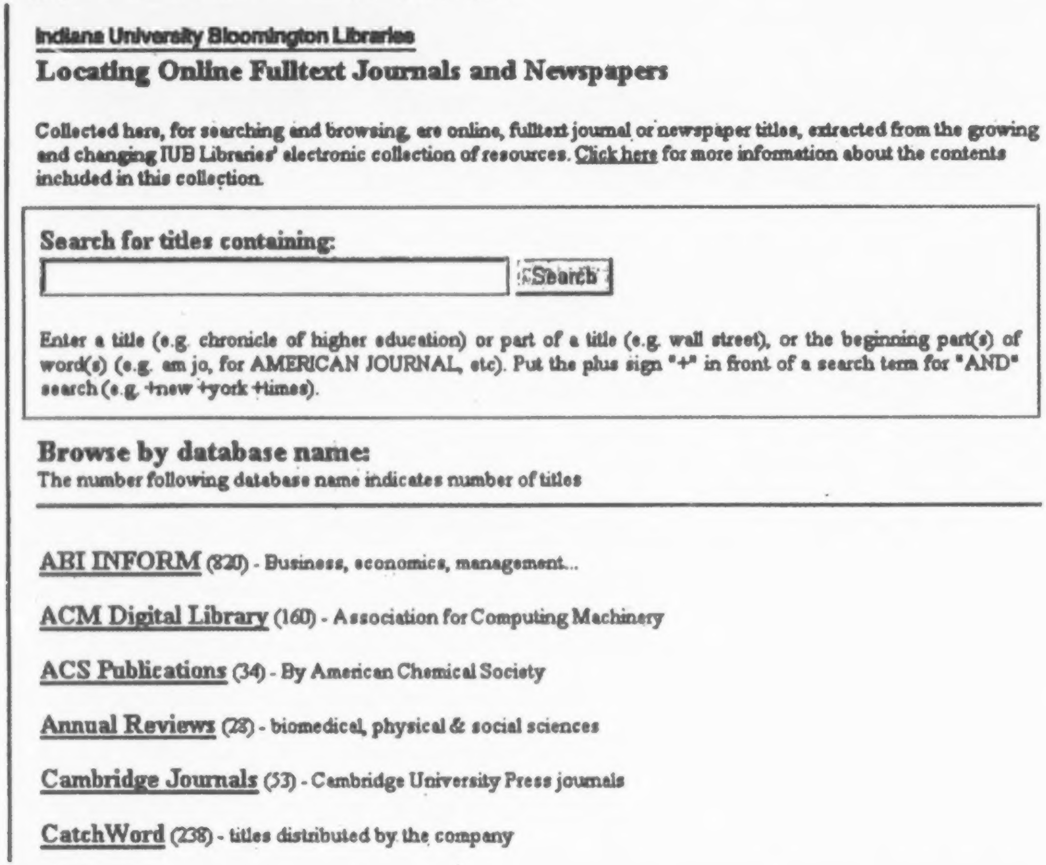
BEST COPY AVAILABLE

The Journal of American History

Publisher: Organization of American Historians

Moving Wall: 5 years

In 1964 the *Mississippi Valley Historical Review*, published by the Organization of American Historians, became *The Journal of American History*. The change in title reflected not only an awareness of a growing national membership in the Association, but recognized a decided shift in contributor emphasis from regional to nationally-oriented history. *The Journal of American History* remains the leading scholarly publication and journal of record in the field of American history and is well known as the major resource for the study, investigation, and teaching of our country's heritage. Published quarterly in March, June, September and December, the *Journal* continues its distinguished career by publishing prize-winning and widely reprinted articles on American history. Each volume contains interpretive essays on all aspects of American history, plus reviews of books, films, movies, television programs, museum exhibits and resource guides, as well as microform, oral history, archive and manuscript collections, bibliographies of scholarship contained in recent scholarly periodicals and dissertations.



Indiana University Bloomington Libraries

Locating Online Fulltext Journals and Newspapers

Collected here, for searching and browsing, are online, fulltext journal or newspaper titles, extracted from the growing and changing IUB Libraries' electronic collection of resources. Click here for more information about the contents included in this collection.

Search for titles containing:

Enter a title (e.g. chronicle of higher education) or part of a title (e.g. wall street), or the beginning part(s) of word(s) (e.g. am jo, for AMERICAN JOURNAL, etc). Put the plus sign "+" in front of a search term for "AND" search (e.g. +new +york +times).

Browse by database name:

The number following database name indicates number of titles

ABI INFORM (820) - Business, economics, management...

ACM Digital Library (160) - Association for Computing Machinery

ACS Publications (34) - By American Chemical Society

Annual Reviews (28) - biomedical, physical & social sciences

Cambridge Journals (53) - Cambridge University Press journals

CatchWord (238) - titles distributed by the company

Part and parcel of this is how we use this data to facilitate generalized user access to these resources. I mentioned my feeling that a collection-level record for electronic sets has little utility beyond serving as a locus to record payments. I think that most technical services and collection development librarians agree with me in this regard. And, while some libraries have extended the scope of this collection-level record by including in it so-called analytical added entries for individual serial titles, the extent of most aggregations renders that solution impractical, if not impossible. Furthermore, communicating a record for resource sharing that might be enhanced or, better, burdened, with hundreds or even thousands of analytic titles is beyond the scope of the MARC record. Lastly, let us not forget that a title entry within a collection-level record is still bereft of all the other access points such as subjects and responsible corporate bodies that our users have every right to expect.

469 Even if it were possible, however, such a listing has minimal utility unless it resides in a Web-enabled, clickable catalog. Thankfully, most of us are finally reaching that point in our integrated library systems development. In the meantime, however, most libraries of which I am aware have taken the approach of creating Web listings of Internet resources. For the most part these include the major database offerings, such as Lexis-Nexis, JStor, etc., as well as an itemized list of electronic journals, both for individually obtained ejournals and sometimes for those included within aggregated databases. For the latter, you will typically find this where it has proven feasible to create a list either because the number of journals comprised in the database is manageable or is facilitated by a publisher-supplied inventory.

470

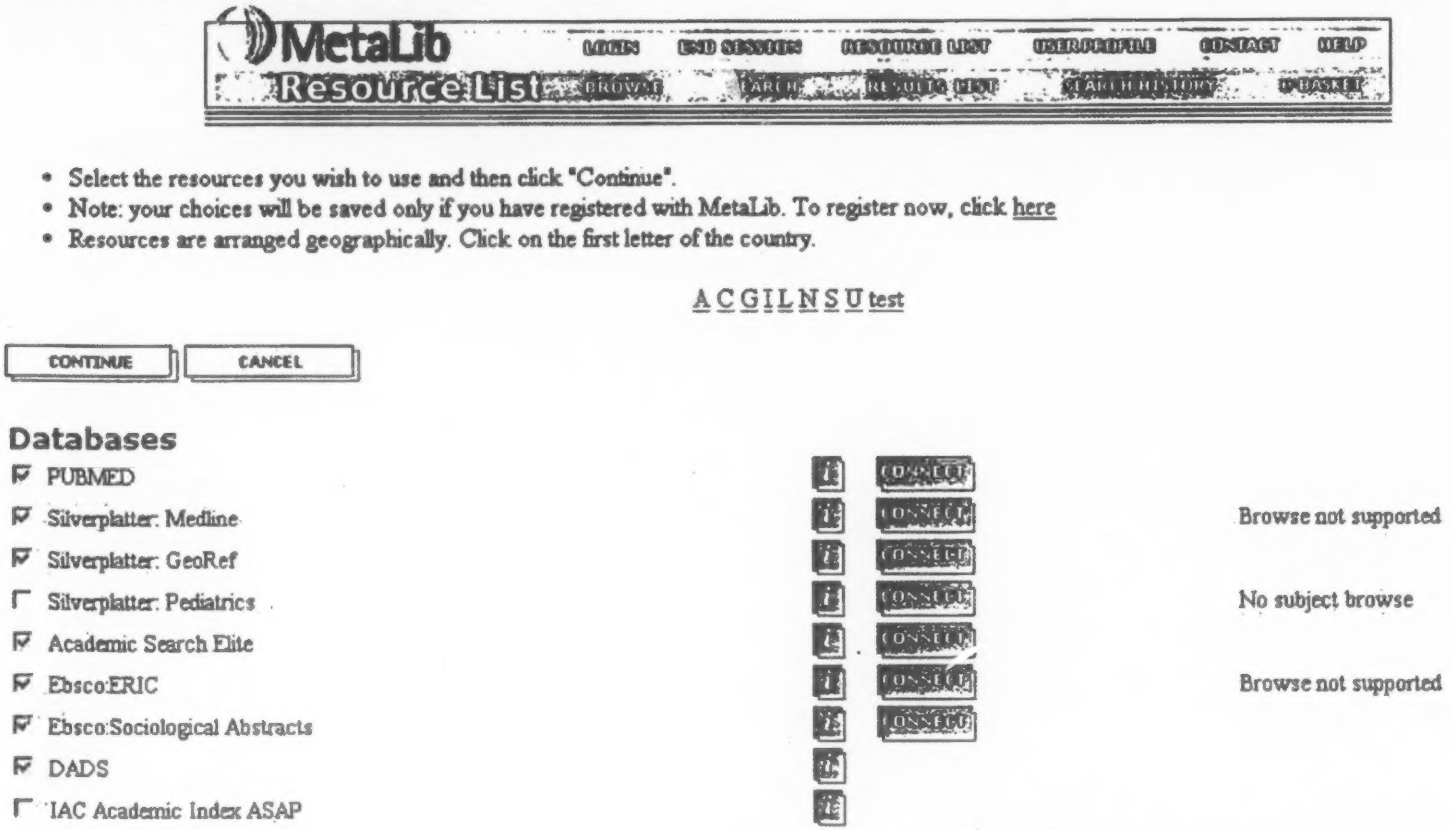
So, then, libraries have created various work arounds, most involving Web listings of their electronic journals. This is an unfortunate but largely unavoidable consequence of technological change coming to the library in varying stages. It greatly troubles me because it represents duplicate, wasted effort at a time of diminished resources and because it takes effort away from the centrality of the catalog. One might argue that in an age of digital collections the catalog is destined to lose its centrality anyway. So why should I worry?

To this argument I have two objections. The first is that I distinguish between textual journals, whether paper or electronic, whether born-digital or digitally reincarnated, and other sorts of digital resources that live on the Internet. Moreover, I believe that our users do so as well. Secondly, I am emboldened by a survey that Indiana University-Bloomington conducted recently of its users' behavior regarding ejournals in which they discovered that their users prefer search and discovery that begins with the catalog, not with a Web page listing. I feel this assertion is corroborated by plans that Pat Sabosik and Daviess Menefee of Elsevier Science Direct made at a conference sponsored by Der Kooperative Bibliotheksverbund Berlin-Brandenburg (KOBV) in June 2000. (As a matter of product placement, I should note that KOBV is an Ex Libris' ALEPH 500 consortium.) It is their belief that local systems will serve as the primary information portal for ejournals and that more types of information and linking will begin to occur within the local systems context.

The information paradigm of the future will soon resemble this flow:

```
* Discover==>Navigate==>Locate document==>
Order document==>Obtain document (Document delivery)
```

And what of digital resources not under bibliographic control in the catalog proper but still somehow listed as part of a local or remote digital collection? How will the users access them? My fond expectation is that we are beginning to see the development of search interfaces and search engines that will simultaneously search and unify heterogeneous databases consisting of MARC records in all the bibliographic formats, as well as databases comprising all present and future data structures: Dublin Core, EAD, TEI, CIMI, VRA Core, MAB, MAB2, etc. This is the promise of the new MetaLib(tm) product already being tested by Ex Libris in several European libraries. It is a recognition that one standard will no longer rule the roost and that we have great need of a Universal Gateway for our users. It can or will handle many of the data format and structure issues, but by itself it will not handle different authority conventions. We need to adopt Barbara Tillett's approach to do that.



- Select the resources you wish to use and then click "Continue".
- Note: your choices will be saved only if you have registered with MetaLib. To register now, click here
- Resources are arranged geographically. Click on the first letter of the country.

A C G I L N S U test

Databases

☑ PUBMED	
☑ Silverplatter: Medline	Browse not supported
☑ Silverplatter: GeoRef	
☐ Silverplatter: Pediatrics	No subject browse
☑ Academic Search Elite	
☑ Ebsco:ERIC	Browse not supported
☑ Ebsco:Sociological Abstracts	
☑ DADS	
☐ IAC Academic Index ASAP	

By channeling our bibliographic energies in this direction we can maintain the integrity of our traditional authority controlled databases and expand our concepts of bibliographic control in new directions. It will not be easy to accomplish this, and not just because of the requirements of crosswalks and data mappings, but rather precisely because of the difficulties of conjoining files subject to different versions of authority control or non-control.

There is one further aspect to all this, too, and that is user preference and user understanding. In general, I have maintained that a Mulver record best serves users provided that the display can be made clear and sensible. That means that all the information that is pertinent only to the reproduction needs to be clearly associated with the reproduction. Problems still abound for issues of searching and proper indexing where format or date of reproduction issues are concerned. (Though I would hope that dual indexing of 008 and 006 format types and somehow capturing and indexing original date and date of reproduction be factored into the equation as well.) Still, as a user in a long ago graduate lifetime, I will say that finding all the relevant records, regardless of format, in a single, intelligible bibliographic construct would have been the arrangement I would have found most appealing. As I have already commented, it has advantages and disadvantages for technical services, and the short-term solution many have opted for may not in fact be the best long-term solution if we can get content providers and system vendors to cooperate on a shared approach to the problem.

- My second observation concerns ancillary or adjunct data.

BEST COPY AVAILABLE

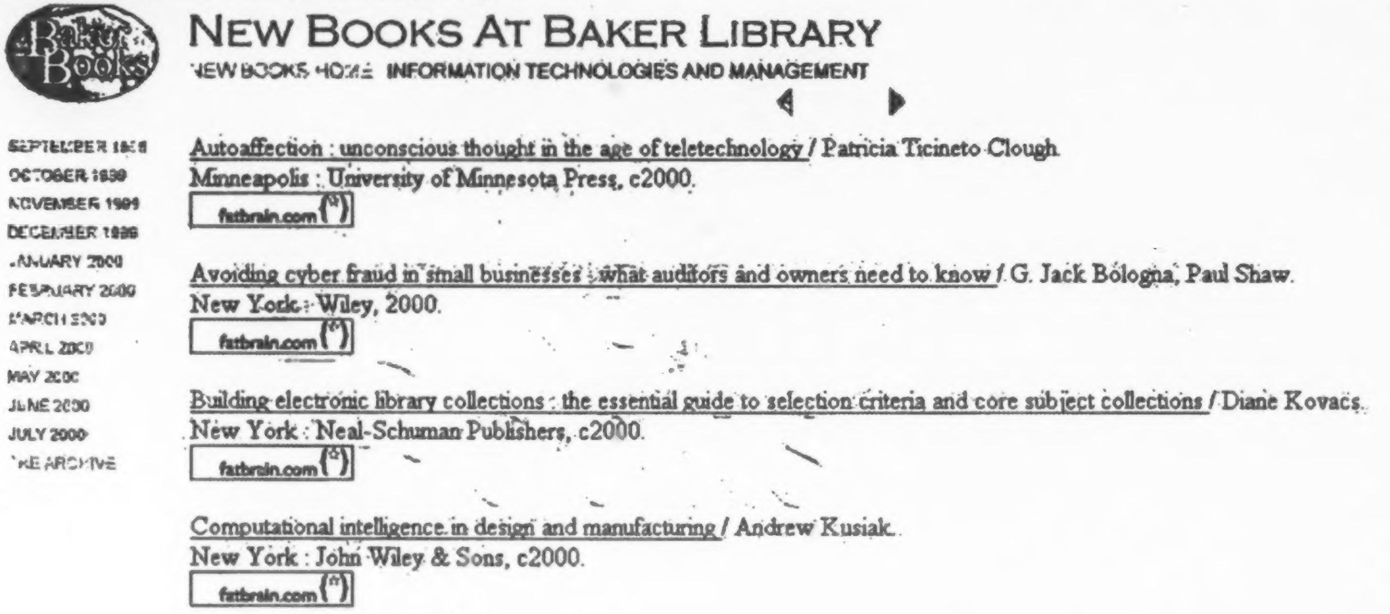
To start with, you must understand that I do not mean 'ancillary' or 'adjunct' in a demeaning or derogatory fashion. Among the oldest examples that come to my mind are A&I databases and table of contents, such as those pioneered by Blackwell North America and others in the late 1980s and early 1990s. For tables of contents, a library would subscribe to some or all of a set, or purchase TOC data for individual titles. Local systems have developed the means to load individual tables of contents records into the actual OPAC where the TOC data is generally keyword searchable. Along with similar sorts of data, such as dust jacket blurbs or back of the book data, there is a wealth of useful data here for our users. Tables of contents often offer a contents-rich vocabulary on the chapter level that goes far beyond the limitations of our controlled subject terminologies and therefore has great appeal to our users. Moreover, while our copy-cataloging staffs who labor in a manual environment have long been held hostage to limitations on the length of those TOCs, scanning technology and, even better, electronic texts offer elegant solutions to the problem posed by manual keying. A few years ago the inventive and technologically adept wizards at the Library of Congress embarked on twin attempts to enrich their records with TOC data through their electronic CIP and business reference books experiments. In a sign of the dynamic future of the record that I am envisioning, their business project did not directly and physically encapsulate this information within the record, but rather pointed to it with a URL. (Of course, this is also one way to overcome some of the limitation of length in the MARC record.)

TOC data is but one piece of a constellation in a galaxy of similar constellations: why do we not add back of the book indexes, author portraits, author pictures, fly-leaf information, back cover information, summaries, or book reviews? Along with TOC data, these are parts of my dynamically enhanced record. To represent that perspective, I have invited Jeff Calcagno to join us from Syndetic Solutions, Inc. Syndetic Solutions, whose motto is "Bringing books and readers together", is in the catalog enrichment services business. In addition to tables of contents, they provide fiction and biographical descriptors, cover images, author notes, and summaries and reviews. I know from my brief stint at Indiana that reviews were a hot topic among the CIC libraries, and that and TOC data is how I first came to learn of Syndetic Solutions. Among the challenges here is that of linking up TOC data held in a central repository with a database that will have both its retrospective and ongoing aspects and adding that magical TOC button-or the magical added info button-to let the user know more information is available. Beyond that there is the technical challenge of continuing to provide keyword access to the valuable TOC metadata if the TOC metadata is not actually already resident in the catalog but only retrieved when requested.

- Third, there is the challenge posed by the rapidly growing body of full-text digital books and providing appropriate metadata for them.

Metadata for electronic books is the key to discovering and ordering from vendors such as netLibrary. This is data that may reside in our catalogs, but even more appropriately on the Net, and it should appeal not only to librarians, but also even more directly to the end user. The data might be, and often will be, based on that for the paper analog of this digital document. It might be available separately or as header data for the text file. These vendor databases might be populated with typical publisher information, or they might be replete with full catalog data, enhanced in their own turn by adjunct data. Indeed, Amazon was the first place that I encountered the concept of online book reviews intimately associated with the book ordering process. But as we in libraries begin to ponder the prospect of purchasing and loading records for thousands of electronic books from vendors such as netLibrary, we need to ensure that we have available, once again, the proper metadata for bibliographic control and access.

I am delighted, therefore, that Lynn Connaway, Vice President for Research and Development of netLibrary, is here today to talk to us about the role of metadata in their strategic plans.



Baker Books

NEW BOOKS AT BAKER LIBRARY

NEW BOOKS HOME INFORMATION TECHNOLOGIES AND MANAGEMENT

SEPTEMBER 1999
OCTOBER 1999
NOVEMBER 1999
DECEMBER 1999
JANUARY 2000
FEBRUARY 2000
MARCH 2000
APRIL 2000
MAY 2000
JUNE 2000
JULY 2000
THE ARCHIVE

Autoaffection : unconscious thought in the age of teletechnology / Patricia Ticineto Clough.
Minneapolis : University of Minnesota Press, c2000.
fatbrain.com (*)

Avoiding cyber fraud in small businesses : what auditors and owners need to know / G. Jack Bologna, Paul Shaw.
New York : Wiley, 2000.
fatbrain.com (*)

Building electronic library collections : the essential guide to selection criteria and core subject collections / Diane Kovacs.
New York : Neal-Schuman Publishers, c2000.
fatbrain.com (*)

Computational intelligence in design and manufacturing / Andrew Kusiak.
New York : John Wiley & Sons, c2000.
fatbrain.com (*)

In my future paradigm that starts with Discovery and ends with Ordering and Document delivery, inclusion of titles such as those netLibrary and others provides a crucial link. The library and the library catalog become conduits not only for discovery and for the full-text, but potentially also for the user to obtain personal copies. Needless to say, this is not limited to delivery of electronically available documents. The Harvard Business School's Baker Library is an example of a library that showcases its new books on its electronic bookshelf and makes available direct links for ordering. This is an example of non-traditional document delivery-one where the user pays for the privilege of having his or her own printed copy and having it delivered direct to the doorstep-and it is also an example of institutional entrepreneurship where the library receives a percentage of the purchase price for funneling the purchaser to the online bookstore.

- I have one final, fourth bibliographic challenge. This is the world of the Internet at large.

This is a universe that is largely uncontrolled and, I dare say, uncontrollable except insofar as Internet search engines such as AltaVista or Excite or Lycos can make any claim to indexing and retrieval. But, even within this vast and chaotic jungle where a new Internet domain is registered every 3 seconds day and night, there are pockets of rationality and hope. I say 'pockets' because I think it too much to think that we could or would even want to catalog the Internet. Such an attempt would be not only doomed to failure, it would be a foolish exercise in cataloging an electronic rubbish heap. We used to have a needlepoint sign in the Office for Systems Planning and Research at Harvard University: "The DUC (i.e., The Distributable Union Catalog) is not a Dump." That referred primarily to the quality of the catalog records, but today we could as easily say that it refers to the quality of the resources for which we want to provide quality metadata. We do not want to consume our precious and limited talents on individual homepages or on junk-like resources.

So, with this in mind, I am very attracted by the assertion of OCLC's Office of Research that controlling 100,000 top-level, intellectually viable and valuable Web sites in a cooperative venture built on human investments would provide a solid and proportionate investment in control of the Internet. That having been said, I still have to admit a residual attraction in long espoused notions that those who create meaningful Internet resources-you will have to determine for yourselves precisely what 'meaningful' connotes to you-should be offered an appropriate, simplified template for metadata self-creation. At least for data archives such as preprint servers and similar repositories I see this as an integral registration service that is the best approach to creating the item-level metadata records that are so important in Internet discovery.

I do not intend to do more than mention non-textual Internet realms, such as those of art images, music, etc. They require more expertise than I can bring to bear on them, but they, too, require metadata, obviously even more than their textual counterparts since otherwise they are otherwise basically un-index-able.

The bottom line in all this is the need for widespread cooperation and for standards. It is, or at least should be, a fundamental tenet of all cooperative ventures that there be standards embedded in them, and it is likewise fundamental that there be widespread consensus in standards building. There is nothing more frustrating than trying to bring order out of a bibliographic chaos that could have been avoided had appropriate standards been established from the get-go. I cannot help but think of all the hard work we would have made unnecessary had we as a profession adopted Pinyin instead of Wade-Giles as our standard for Chinese Romanization 20 years ago when the body of Chinese bibliographic data was relatively small. Yet, that is where we find ourselves because librarians, specifically technical services librarians, were not fully involved in either the birthing of the Web as a home for Internet resources or in the first attempts at devising a non-cataloging metadata standard. As libraries move with greater assurance into digital collection development both on the creation and collection levels, it will be paramount that we get our bibliographic control house in order.

But that is not and cannot be the entirety of the argument. We need to accept the fact that more materials, both printed and digital, will always demand more of our attention than we can possibly accommodate. As I observed earlier, this was true in a print-only world and it is exponentially true now in the current environment. We need to consider the feasibility of two complementary approaches to this overwhelming bibliographic tsunami.

DigiTool Library

Solutions for Building
Digital Collections
from Ex Libris

www.exlibris-usa.com

BEST COPY AVAILABLE

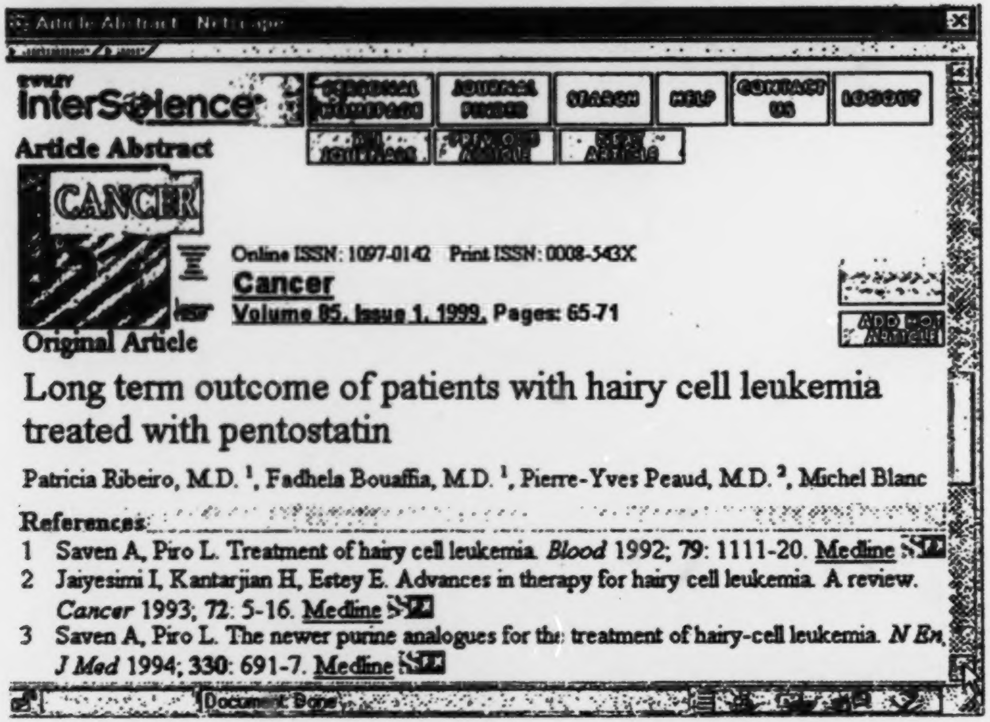
The first is employing more technology, and more sophisticated technology at that. I have in mind machine analysis of digital documents. For PDF documents this will involve OCR conversion as the basis for full-text indexing and then automatic metadata creation based on the OCR derived output. For documents that are already XML-based, as I expect will ultimately be the case for most non-traditional metadata and ultimately even for most MARC data, we will see automatic metadata creation-or perhaps for an indeterminate period of time-machine-assisted, cataloger-approved metadata creation. This is the plan espoused, for instance, by Ex Libris for its upcoming DigiTooLibrary(tm) product. This will follow the model of the mid- to late-1990s authority record creation that has come to be a staple of participants in the NACO program. It has proven itself successful and been a tremendous help in automating the more mundane, predictable parts of authority control.

If we are to see this concept through to implementation, though, more is needed than just standards and technology. What is needed will be true partnerships and true commitment, substantially greater than we have now in OCLC or RLG where librarians often pay lip service to cooperation and to trust. Rather, they maintain lists of acceptable and non-acceptable libraries and are constantly re-inventing the bibliographic record to meet their own, internal and internalized standards. I have said it often before in venues devoted to issues of cataloging of printed materials, and let me say it here as well. We cannot afford such arbitrary distinctions. The days of golden records are long gone. I once heard Sarah Thomas quote Winston Tabb as saying, "What good is a bibliographic record if it is not there when we need it and at a priced we can afford?" This is all too true. If the world is simply not to pass us by as an anachronistic profession, or if we in technical services are not to become an archaic sub-profession, we must adapt and seek out new partnerships. These partnerships will be technology-based and help us to do more faster and more accurately, and they will also be based on new arrangements with content creators, both individual and corporate, and with corporate content providers. These partnerships should be based on mutual understandings of what is desirable and what is possible, and a realization that not all that is desirable in terms of traditional bibliographic control is within our grasp even with the most advanced technology.

OCLC's CORC project is an attempt in this direction. I have been watching the development of the software for about 2 years now. It has the right elements in its repertoire, it has the correct pieces of an Internet toolkit, and OCLC has spent millions of dollars of research time on it. Is it all that we want or hope? No, at least not yet. But it does show that metadata about the new Internet frontier is crucial to our collective future well being.

I was recently reminded of a pithy aphorism that I formerly used in talking about standards and workflows: "Better is the enemy of good enough". Regina Reynolds, of the National Serials Data Program, put a name to my saying when she quoted Voltaire as saying, "The perfect is the enemy of the good." My saying is a slight variation on Voltaire, to be sure, and I had no particular source for it. I have to admit, with some sense of embarrassment, that it does not sound particularly self-flattering to admit that you are willing to settle for less than perfection. To return to Matthew Beacom's comment about NaīveMetadata, the fact is that searching and linking that is dependent on metadata can only be as complete or as precise as the metadata itself is complete or precise. But, in telling my former staff at Indiana that there is the Good and the GoodEnough, I also pointed out that metadata description, taken in its full context of description and classification, is an art, not a science, and certainly not an exact, reproducible science. Those of us who have long considered ourselves the guardians of the bibliographic universe need to broaden our horizons and recognize that the world is not as narrow as we once defined it. It requires new approaches, new technologies, and new allies. Let us accept that and frame the debate in wider, more inclusive terms than formerly.

I do not want to end on a 'down' note. The subject is too important and the consequences of failure or failure to act are too high. So I want to draw your attention to some recent research and a new opportunity to link together our disparate metadata resources. This comes in the form of a relatively new adjunct to the Net and the Web, and I became aware of it, coincidentally, as I left almost 30 years in academia for the vendor world. This is the technology created by Herbert van de Sompel and his colleagues at the University of Ghent. Called SFX, for Special Effects-and not to be confused with the SFX technology that exists primarily to deliver audio-visual resources over the Web-this is nothing short of a revolution in how we should envision research on the Web. The University of Ghent is an Aleph 500 site, so it is no surprise that Ex Libris, the creator of the Aleph library management system, saw the potential in this product and now has an exclusive right to license this technology. SFX is a framework for context-sensitive linking between Web resources. It is the means to unite or link disparate, heterogeneous electronic resources such as abstracts and full-text, all the while keeping in mind the context in which the user works and that some sources of data may be institutionally more appropriate for that user than others. It also has the ability to link to related occurrences of authors, subjects, and other metadata access points.



Article Abstract

Online ISSN: 1097-0142 Print ISSN: 0008-543X

Cancer

Volume 85, Issue 1, 1999, Pages: 65-71

Original Article

Long term outcome of patients with hairy cell leukemia treated with pentostatin

Patricia Ribeiro, M.D. [1], Fadhela Bouaffia, M.D. [1], Pierre-Yves Peaud, M.D. [2], Michel Blanc

References

1 Saven A, Piro L. Treatment of hairy cell leukemia. *Blood* 1992; 79: 1111-20. Medline
2 Jaiyesimi I, Kantarjian H, Estey E. Advances in therapy for hairy cell leukemia. A review. *Cancer* 1993; 72: 5-16. Medline
3 Saven A, Piro L. The newer purine analogues for the treatment of hairy-cell leukemia. *N En, J Med* 1994; 330: 691-7. Medline

This is truly exciting stuff. Yet I am struck by the notion, heard all too often but I do believe true in this case, that we have hit on one of the Holy Grails of research. The Holy Grail is that of 'seamless interconnectivity'. To back up a step, though, this technology is seamless only because the metadata exists as seams in an information architecture. SFX then takes the seams one step further and turns them into a library-defined seamless whole. Please note, that is 'whole' and not 'hole'.

What is most gratifying about the SFX solution to reference-linking is its unabashed reliance on metadata. The fuller and more accurate it is, the better the reference-linking that will result. In a world with result sets numbered in the 1,000s or 10,000s, precision and recall are constantly at odds. I have therefore found myself intrigued by the concept of metadata as the 'wrapper' of Internet resources. Now, it is true that wrapper has a specific meaning beyond what I am according it here. To me, if I can make use of a bit of license, the metadata wrapper is as simple as that of the gift wrap of a present or of a candy bar wrapper. In the case of a present, the purpose of the gift-wrapping is to disguise or even hide the contents. In the case of a candy bar, the wrapper serves to convey all the information required to accurately know what is contained inside: the brand of candy bar (= title), the manufacturer and place of manufacture (= imprint), the weight (= collation). It also has a key number (scannable barcode), list of ingredients (= table of contents?) and nutritional information (based on a 2,000 calorie/day diet), and the notation that it is Kosher-Dairy. So, for example, my half-serious attempt at cataloging a Nestle(r) Crunch(r) candy bar:

BEST COPY AVAILABLE

Nestle(r) Crunch(r) [candy bar].-Glendale, CA, USA : Nestle USA, Inc., Confections Division, [2000].
1.55 oz.
"28000-13170."
"OU-D."
Nutrition facts: Calories, 230; Fat cal., 100.
Ingredients: milk chocolate ...

This is important metadata because not only does it ensure that I do not buy a bag of M&Ms(r) when what I really want is a Crunch(r) bar, but it gives me all the essentials except the price. Content providers could do worse than imitate Nestle(r) Inc. What I do not want them to do is give me an information present enclosed in metadata gift wrapping that gives no clue to the informationpresent's true identity or, even, worse, to give me the information present with no metadata wrapper at all. Think of it as an anonymous candy bar or a tin can that has lost its label. Who would have any interest in such an object?

Those of us who are devotees of and believers in bibliographic control need to recognize the absolutely pivotal role metadata has in the new information economy. Many times in recent months I have heard that we need to claim our rightful place as information managers. Morever, the truth is that the information we now must manage is of an entirely different magnitude than what we faced before. If we ever thought that we could manage it alone-and I think we were mistaken if we thought we could-the fact is that we can no longer do so. We need to seek out and develop our natural partnerships in the information and systems communities and make certain that these partnerships are based on a shared vision.

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Vendor Partnerships For Bibliographic Control

Amira Aaron
Director, Marketing and Programs
Faxon, RoweCom Library Services

Panelist, November 16, 2000
Library of Congress Bicentennial Conference on Bibliographic Control for the New Millenium

Final version

Good afternoon. Michael Kaplan asked me to speak a little about the aggregator challenge in regards to bibliographic control of electronic resources. I will cover that briefly, but I also want to add a few other comments about the more general topic of this conference - from the view of a librarian working for a vendor, but with the emphasis on "librarian". When I speak about vendors, it will be in the more generic sense, including ILS and other vendors as well. I will also touch on the role of our publisher colleagues.

I'd like to start by briefly reviewing a few underlying assumptions for my comments, many of which we have heard at this Conference. Unquestionably, new business models for acquisition and access will require different levels and sources of cataloging, or metadata. Articles, websites, pre-print services - all of these will need controlled, consistent integrated access. Libraries cannot do it all and must select where to focus their resources. Practical, scalable solutions need to be found, such as those outlined by Regina Reynolds in her excellent paper. Michael Kaplan is on target with his concept of a core bibliographic record enriched with contributed data from a number of sources, including publishers and vendors. And there is no question that libraries can, and must, benefit from appropriate partnerships, with publishers, vendors, search engines, Amazon.com, and many other commercial enterprises.

So what are the specific issues surrounding aggregators and bibliographic control? Aggregations have proven to be a cost-effective method of providing widespread access to the full text of e-journals. Although there has also been widespread concern about bundling together a large number of titles that haven't been "selected" in the traditional sense, some recent studies, including one at OhioLink, are showing that the titles that were not previously selected for a library's collection are receiving as much use, if not more, than the previously selected titles. So aggregations are very much here to stay.

The nature and size of aggregations can vary widely, in content, coverage and business model, so that it is difficult to deal with a uniform set of processes or standards for bibliographic control of these collections. One title can be part of multiple aggregations, which only exacerbates the multiple version problem. Titles move in and out of aggregations, often without notice to the aggregators themselves. Libraries want to integrate access to titles in aggregations with their other resources and patrons need to be able to link to the "appropriate copy" of the title to which they have authorized access.

Ideally, records for the individual titles should be created once and used by many; we need records for both the titles and electronic holdings. These records also need to be maintained, updated, deleted, etc. And, regardless of the source of the record, we need standards and some agreed upon level of quality enforcement.

Current methods of handling access to aggregator titles vary widely from library to library. Some offer no access to the individual titles, but simply a record for the aggregation itself in the OPAC - hardly a satisfactory solution. Others offer a web page link to the aggregation. More often, we find a web page link from the e-journal title to the aggregation. Links to the Jake project at Yale are becoming more prevalent - from either the OPAC or webpage, or both. For those libraries offering access to individual e-journal titles from the OPAC, there are sometimes multiple records for each version or aggregation for access. Others use a single record containing holdings for multiple versions. The publisher URL (or a durable URL) is displayed either in the bibliographic portion of the record or with the appropriate holdings information (this latter approach is much clearer to the user and hopefully will become more widespread).

There are also multiple sources of records for aggregated titles currently being used in libraries. Some are locally cataloged and maintained, although it is generally agreed that this is not a scalable solution. There are also cataloged sets from OCLC contributed by participating libraries. Some consortia offer cataloging records for titles held jointly, such as the NESLI group in the UK. Sometimes vendor lists from websites are run through a MARC program on a regular basis and loaded into the ILS. Others use Jake records downloaded in MARC format. Then there are aggregators who are able to offer MARC records, either themselves or by using a commercial MARC service.

What are some of the challenges faced by aggregators in attempting to provide records for their collections? Libraries who purchase one aggregation may not all have access to the same set of titles; often there are different packages within aggregated sets. In the case of RoweCom's Information Quest, for instance, access is provided only to those titles for which the library has a licensed subscription with the publisher. So one solution doesn't fit all. Processing bibliographic records routinely and mechanically for aggregator titles has its pitfalls. Aggregators don't as a rule examine an electronic title to determine if this is simply an electronic version of a print title or if there is in fact new content. Sometimes there is confusion about whether there is actually full text or the site simply contains abstracts.

The aggregator may need to create a MARC record if no print equivalent exists. Higher-level staff and training is often necessary. I would actually support an option for the aggregator to "outsource" the cataloging back to libraries for payment. In this way, both parties would benefit. The process of creating bibliographic records needs to be cost-effective for the aggregator. Be assured that the cost will be passed on, bundled or not. And indeed the library community does need to pay fairly for added value services.

There are other difficulties inherent in relying on the vendor for bibliographic records for their collections. Often the management and priorities at the vendor change and resources are no longer available for the cataloging project. Aggregators need improved monitoring to know when a publisher or title drops off or changes, or the format specifications for issues change. They need a way to easily maintain the data - deletions, holdings, coverage, etc. - and to do it centrally and consistently regardless of which ILS system is involved at a particular site.

Many of you are probably familiar with the work of the PCC Standing Committee on Automation Task Group on Journals in Aggregator Databases, so I won't cover this in depth. Following a CONSER survey which demonstrated that the majority of respondents wanted vendor-supplied cataloging records for electronic titles in aggregator sets available in the OPAC, the Task Group was charged with proposing the content of a vendor-supplied record for an "aggregator analytic" and mounting a demonstration project. They were also to make recommendations for maintenance and updating. A final report was issued in January 2000 and this contained practical solutions for the issue at hand. A new task group has been formed to continue the work, dealing with record sets, e-books, communication, increased work with vendors, and more.

The PCC Task Group determined that the type of record should depend on the number of titles in an aggregator database. For a very small number of records, human-created analytics were best in terms of quality, but not scalable for larger collections. The second-best solution consisted of machine-derived analytics from the print version of the serial and assumes the availability of necessary cataloging records. Beyond 200-300 titles, the machine-derived solution was selected as the best option. Other choices considered were machine-generated analytics which rely on defaults, scripted creation of minimal records, and a single combined coverage index, like Jake. Although some vendor-supplied aggregator records have been available and more are becoming available, their use is still disappointingly minimal. It is hoped that this volume will increase in the future.

It appears that we need to come up with a more central and granular solution for these records. Michael

Kaplan is correct in wishing for an EDI-like solution for vendors to update holdings and URL's on a timely and straightforward basis. We need to do this once in a central database and then have the holding libraries notified about changes, or find a standardized way to send automatic updates to all ILS systems to update selected portions of the record. We have been successful in loading and updating EDI invoices; if we can do it and trust the process where money is involved, surely we can come up with consistent match points and a process to update holdings and URL's. We also need to increase the level of granularity of OPAC access; one ILS vendor has shown interest in receiving from us a SICI-like string including a durable URL - to create electronic holdings which can then link to the table of contents at the issue level of an electronic title. I would recommend that we need a SISAC-like group including librarians, ILS vendors, utilities, aggregators and publishers - to find solutions and work together to implement them quickly.

Let's take a moment to look briefly at some issues surrounding publisher responsibility for metadata in the future. For the first time, publishers are realizing the intrinsic value of their metadata in relation to e-commerce applications and will be more interested in solutions which lead to increased sales of their publications. The results of bad data and errors will be more readily apparent and have a negative impact on sales. So we have an important opportunity here, as publishers will need to begin collecting metadata in a more standardized form from their authors. We should actively participate with publishers to ensure that they will be distributing this metadata for titles, articles, chapters and related names and works in a standardized and consistent format.

Publishers should have increased responsibility as well for the quality, accuracy and updating of bibliographic and other metadata. They should be informing us immediately or before the fact about title changes and holdings coverage changes. And publishers should be partnering with libraries and vendors to ensure consistency and quality of their data. Use of library authority files and authority processing would benefit the publisher and the user of all online services. Publishers are also providing increasingly enriched data - tables of contents, resource links, author biographies, issue dispatch data, rights management information, etc. Libraries need to be actively involved in the standards, such as ONIX, which are being developed to deal with these new types of metadata. This is one particular partnership with standard groups and publishers that should immediately be explored.

What are some of the vendor roles in creating and dealing with metadata for electronic resources? Vendors should be creating the umbrella systems for resource discovery, integrating both local and networked resources. And they should be doing this with much library input. They should be developing and applying technological solutions for bibliographic control and record enrichment. Vendors should be actively partnering with library, publisher and other groups/vendors in standards development. They need to be encouraging and publicizing the use of library defined standards by publishers and authors. When appropriate, vendors should work to provide and maintain standardized metadata (cataloging data) for their collections. And they should be providing enriched data and links from the standard bibliographic record.

In preparation for our topical breakout sessions, I wanted to see if there were some lessons that we as librarians could learn from the commercial sector and keep in mind while carrying out this daunting task! This is the list I've come up with:

Competition

For the first time, libraries are facing serious competition in their traditional functions and areas of expertise. Unfortunately, the Internet is now the first place that many audiences turn to for research; libraries do not yet have an obvious place on the Net nor are they the first place that the average user now thinks of for information access. It is time to actively work to regain and retain our market share! If we have to borrow some tactics from our "competitors" to be successful, so be it!

Partnership

The commercial world is now creating partnerships left and right; companies can no longer go it alone. Yesterday's competitor is today's strategic partner.

Cost/Benefit Analysis

The commercial sector constantly performs cost/benefit analyses. And we librarians need to do that as well. We have to choose what's important and identify those areas which are less important and where perfection is not in fact necessary. I was struck yesterday by Barbara Tillett's talk on the new possibilities for authority control and thinking that if we could be successful in working with publishers and search engines to implement some of these features, this would be really significant - in my mind, much more significant than worrying about exact transcription or correcting someone else's cataloging copy. Something needs to give - let's concentrate on where we can do the most good and have the most positive impact.

Marketing
As has been said multiple times in different ways during the last couple of days, libraries need to learn how to better market themselves and their knowledge and skills! We must be assertive and prove our value add in as concrete ways as possible. We have so much expertise in resource evaluation, authority control, cataloging, access! At the Charleston Conference recently, we heard that the new ONIX standards were being developed with little or no involvement from the library community. Sitting in a roomful of librarians, I couldn't believe that no one was angry enough to stand up and ask why... We must demand to be heard and to be involved!

Risk Taking
We need to take risks and experiment- it's not a matter of life and death. If a project doesn't work out well, so be it! Some will work and we'll be the better for it! We no longer have the luxury of planning out every last detail and ensuring that an idea won't fail along the way... And we need to provide or secure funding for these experiments, as Jane Greenberg has said earlier today.

Forecasting
Forecasting - this is a hard one - we need to try to predict the future and to look ahead as much as we look at current problems. We need to anticipate future challenges and design our solutions to be flexible enough to meet future needs. Good luck to us all!

Time To Market!
And, finally, we need to be concerned about time to market! The world won't wait. And we don't want to be bypassed because we can't make quick decisions or because we're perceived as bogging down a process.

The library community needs to formulate solid, practical, immediate action plans which include partnerships with the commercial sector, in order to deal with the many challenges facing us and put them into motion. Thank you.

Library of Congress
December 19, 2000
Comments: lcweb@loc.gov

Catalog Enrichment Services
Syndetic Solutions, Inc.

Jeff Calcagno, Director of Sales & Customer Support
Library of Congress Conference
on Bibliographic Control in the New Millennium
16 November 2000

Final version

I want to first acknowledge that many of the bibliographic enrichment data elements which I will be discussing are not new to the library community and library users. For some time it has been well established that the use of tables of contents, summaries, annotations, analytical notes, etc., are a valuable addition to the library catalog and the library users' information-seeking experience. Many individuals at this conference have done extensive research, dating back over twenty years, clearly demonstrating the usefulness of this data.

I should also go on record to state that libraries have correctly responded to these needs by establishing standards, both under cataloging rules and in the MARC format, to make many of these data elements available for their patrons' use, whenever possible. There is no group more qualified to create such information. Of course, finding the time and financial resources to create the data locally has proved to be increasingly difficult.

"Raised Expectations"

Though increasingly costly to produce at the local level, Syndetic's believes bibliographic enrichment data will play an important role in the future development of the library OPAC, library web sites, and what is often now being called the "library portal". If we anticipate leisure reading to continue to increase, and as a large legion of life-long learners march into "retirement", coming to the library, either physically or "virtually", to find a good book is going to take on a much more complex meaning. And, of course, researchers, both non-professionals and those in academia, are also gleaning much more information from sources formerly inaccessible prior to the Web. They now utilize an astonishing array of highly structured abstracting and indexing files and full-text databases unavailable until a few short years ago.

I should also hope none of us are too surprised if a whole new group of users begin to discover all that the library can provide. Many of them are "lining up" now, if you will, at the local online bookstore, and they are experiencing a plethora of information about books in the form of cover images, summaries, annotations, tables of contents, reviews, author interviews and biographies and, of course, the ubiquitous reader reviews (everybody has an opinion!). While the approach to developing these types of enhanced bibliographic databases has appeared to be a "more is better" approach, Syndetics believes that the library catalog is a much different access tool that will require a more careful and discerning approach to fulfilling their users information needs. But it is certain that libraries must begin using new and creative methods for bridging access to their collections in a more comprehensive manner. I also should emphasize that the technology and production capacity is now available to begin bridging these "bibliographic gaps" in a timely manner. It is clearly time for libraries to satisfy the raised expectations of their users.

Enrichment data benefits OPAC users in several ways:

- Improves the users ability to locate and evaluate specific titles of interest
- Improves the precision of resource sharing
- Improves access to underutilized portions of the collection

In addition to Syndetics own enrichment creation efforts, large quantities of useful enrichment data are also now available from thousands of sources, including publishers, book wholesalers, review sources, and others. Not surprisingly, many of these data elements are available in a multitude of electronic formats and editions which is another issue that appears to be a hot topic at this Conference. We will certainly be interested in any developments that take place here in this regard.

Syndetic Solutions

Syndetics was founded by a group of librarians and library researchers over two years ago to provide a single source for a wide range of bibliographic information to enhance library public access catalogs. To this end we have established relationships with publishers, book wholesalers, and review sources to make this information available to libraries and booksellers. By becoming a reliable aggregation source, we are also developing relationships within the library community to incorporate enrichment data into library catalogs.

Enrichment Data

What data is available? Syndetic's set of databases include over 1.5 million separate bibliographic enrichment data elements and it is growing at the rate of approximately 5,000 data elements each week or 250,000 elements each year. And much more is to come.

Tables of contents, summaries and annotations are certainly available.

Syndetics can also provide enhanced fiction descriptors to provide readers with considerable precision in finding works of fiction and biographies. This includes precise genre and sub-genre headings, character names and their personal attributes (e.g., gender, ethnicity, occupations, etc.), geographic settings, series and award information, all fully searchable in the library catalog. Author notes and lists of contributors, which provide useful information about an author's educational background and institutional affiliation, can also be supplied for many scholarly titles.

And we also have available a large number of book reviews and first chapters from a number of sources that cover both trade and scholarly materials.

We also offer cover images, book jackets, cover art or whatever you want to call them. They are often visually pleasing to the eye and add graphics to an otherwise text-filled screen. For some of us, they may even have a useful access feature. In fact, we are now working on a program to create keyword descriptors of cover images as searchable data elements. So you may still yet find that cookbook with the green and yellow cover! Syndetic's presently offers three different sizes of cover images, from "thumbnail" to large, and we are now working with libraries to include them in their catalogs. The library catalog will certainly never look the same.

Enrichment Data Attributes

Scope

Syndetics intended coverage includes English-language monographs currently in-print, and we make every attempt to gather as many enrichment data elements as possible about each title. The majority of our enrichment data covers titles published since 1985, however, Syndetics also manages several retrospective projects that yield enrichment data for many out-of-print titles. We are also now beginning to expand our coverage to include non-English language titles including French, German, and Spanish.

Because Syndetic's receives information from a large number of data providers, consolidating and standardizing data formats is a critical component of our services. Information is delivered to Syndetic's in a variety of formats, including MARC, ASCII (text and delimited), HTML, XML, and XML variants and many proprietary formats.

Timeliness

Giving libraries the advantages of a single source for aggregating enrichment data, without also providing timely availability, will often make the data less useful to users. Whether data is received in electronic or print form for conversion, editing and distribution, it is important that this process takes place in a timely manner. Because many libraries order books prior to publication, much of their cataloging data requires enrichment shortly after the book has been ordered. Most information Syndetic's receives is on a set schedule from our providers and conversion work is often accomplished within several hours; editing work is often accomplished within 24-48 hours. This is an area where libraries may wish to carefully examine performance benchmarks from enrichment data suppliers.

Relevance

Accurate and precise enrichment data improves search access. Tables of contents, summaries, enhanced fiction descriptors and, in the near future, indexes and chapter-level bibliographies, are rich in useful keywords. But not all enrichment data is appropriate for all libraries. As an aggregator of catalog enrichment data, Syndetics continuously evaluates and implements editing procedures that retain useful and consistent enrichment data for libraries, carefully considering the costs versus benefits. We welcome input from libraries and their users on the appropriateness of various enrichment data elements and hope additional exposure to enrichment data of all types will lead the library community to some general consensus allowing Syndetics to train our focus accordingly.

Development

Through our own research, and through discussions with librarians, Syndetics has been working to identify other enrichment data elements that we plan to make available to library users in the future. Let me quickly note four of the most significant programs.

> Indexes - This information is specifically noted in Michael's paper and they are certainly worth discussing. Syndetic's has been working on an index conversion project for some time. A concern for us is how or even whether to attempt a consistent format. There are also issues related to the cross-reference structures contained in many indexes and the inclusion of author names in indexes. Both of these authority control concerns are giving us pause to think carefully about what library users will demand. Finally, the sheer size of most indexes will incur considerable conversion and editing costs. We do envision use of machine-readable indexes for selected works in the near future and we have started working with libraries to determine what types of materials will benefit most from having searchable indexes included in their catalogs and in what format. Once library test partners have been identified, we will begin a pilot project.
>
> "Suggested Readings" & bibliographies - Having search access to this type of information will provide one additional research tool for scholars looking for related research work or attempting to locate works by a given researcher. It will also make "browsing" the catalog that much more productive. Syndetics has completed format definitions for the standard data elements; they will be fully parsed and standardized so they can be hyper-linked to the related titles and authors. Syndetic will begin a pilot project in 2001 to initiate the creation of approximately 10,000 bibliographies over a six-month period which will be made available to libraries that wish to participate.
>
> List of tables, illustrations, graphs, etc. - It is clear to us that this type of data, which is often available along with the table of contents, will provide useful access and descriptive information. The critical task at this time is working with libraries and local system vendors to address display and indexing issues for this data.
>
> Author Profiles - This is a what we are calling an authority record "hybrid". The objective being to allow searching on specific kinds of authors with regional affiliations. It will contain such information as place of birth, current residence, areas of genre or subject

expertise, ethnicity or cultural background, occupation, institutional affiliation, awards or honors, etc. Most early testing will involve booksellers, however, we believe libraries will also find a place for expanded author information in their catalogs.

Distribution & Access

Syndetics provides enrichment data directly to libraries and through marketing arrangements with suppliers of bibliographic services, local system vendors, and providers of web-based search software. The continuing growth and development of these arrangements is critical in allowing enrichment data to come into common use and to promote unique and creative uses of these data both in indexing and display among the many OPAC vendors and other possible outlets.

Libraries that receive enrichment data directly from Syndetics have complete control over determining exactly what types of enrichment data they wish to utilize (e.g., only tables of contents, cover images and reviews), in what format the data should be placed (e.g., MARC fields, HTML, XML, etc.), whether enrichment should be performed retrospectively or only on new titles, whether it should be limited to subsets of the collection (e.g., only juvenile materials) and how often enrichment should occur (e.g., weekly, monthly, quarterly, etc.).

While Syndetics is now providing enrichment data for several different types of library systems, mainly through MARC record enrichment, it does appear that we are moving into a transition period. The traditional "vessel" for holding such data for libraries, the MARC format, is demonstrating that its original purpose, as a well-structured bibliographic communications format, does not appear to be the best place for most, if not all, enrichment data.

Record and field size limitations, though not an issue for Syndetics, and probably not even the MARC format itself, are certainly issues with local system vendors and bibliographic utilities. Screen display concerns for viewing a bibliographic citation with enrichment data are an even bigger issue because library users can be faced with the display of a "never-ending" record that contains more data elements than even the most patient users wish to view. As a result, Syndetics is now working with libraries and local system vendors to make these data elements accessible remotely from separate enrichment files which can be linked to a library's bibliographic record and displayed on a "as requested" basis. Presently, two approaches have been identified and put into practice.

Linking field embedded in the MARC record (e.g., 856)

Placing linking fields in MARC records for enrichment is easily accomplished though some local systems presently have constraints on how various "buttons" will allow for displaying enrichment data from a linked file. While effective for viewing enrichment data, this approach also appears to have the disadvantage of not allowing the enrichment data to be searched in many catalogs. Most would agree that this is a serious drawback for many enrichment data elements, particularly tables of contents, annotations, author notes, bibliographies, and indexes. One remedy is to place some of the enrichment data in the MARC record and simply not display it (which most local systems can do) but this is certainly not an elegant solution. This approach appears to us to be a 'transition solution" that will phase out as software advances occur. The second approach portends this coming.

"Umbrella search" of the OPAC and Enrichment Files

"Umbrella search" software is now available which not only will search across multiple electronic files, but will locate, combine and display basic bibliographic information with corresponding enrichment data. Libraries implementing this approach eliminate the need for manipulation of the local catalog record by Syndetics or the library. This allows the catalog record to be, as Michael notes in his paper, the "center of the bibliographic galaxy" for the library while the enrichment data forms various "constellations". This approach also means that Syndetics is able to focus its efforts solely on the process of managing and continuously updating the enrichment files rather than continuously enriching many thousands of library catalogs. This is the more practical approach to making enrichment data available.

We also believe that the utilization of such software is particularly valuable as libraries seek to further refine the user's search experience for both printed and electronic information. This extends from the support of user search profiles ("My Library" concept) to the use of enrichment data to facilitate automated notification of related titles of interest. By utilizing this data in such a manner, we believe that the "tailored" library catalog will become much more of a reality. The ability to access such data remotely in HTML or XML formats will also allow libraries to customize displays through the use of library-defined style sheets (library "branding"?).

Issues for Discussion

While libraries grapple with the considerable task of providing bibliographic control over the ever-expanding galaxy of information found on the Web, they should not lose site of their own Milky Way, the local OPAC, and the collections contained therein which many libraries have spent decades or even centuries building. As we expand both the amount and scope of available enrichment data, Syndetics continues to seek feedback from libraries and library researchers on enrichment usage. In particular:

- What enrichment data elements do libraries wish to consistently incorporate into their library catalogs?
- How and where do libraries wish to utilize these data? (e.g., Use for only portions of the collection? Establish a "hierarchy of use" for certain enrichment data elements?)
- What data elements should be searchable?
- What data elements should be displayed?

Certainly, serious discussions, with the objective of establishing specific guidelines or standards, will assist vendors in responding to the needs of libraries in this regard. This is one of the reasons Syndetic's is pleased to be participating in this important conference.

Syndetics does believe that libraries should aspire to and demand well-crafted, complete and timely bibliographic enrichment data. The attributes of the data must reflect these demands and be integrated in such a manner as to respect the considerable efforts that have been put forth by cataloging staff to maintain the integrity of the library catalog as an access tool for their users. While the online bookstore is often pointed to as a model for libraries to follow when considering the addition of bibliographic enrichment data, the comparisons end quickly when issues such as authority control and the dilution of search relevancy are closely examined.

Providers of enrichment data should bring to the task a considerable amount of experience in handling enrichment data and managing bibliographic files, being particularly aware of, and sensitive to, the many issues related to catalog maintenance which can sometimes be a source of conflict between the technical services and public services staff. However, we also hope that vendors and libraries are willing to experiment with enrichment data in unique and creative ways to help make the library catalog or "library portal" a more dynamic and effective information-seeking tool for their users. Syndetics welcomes the opportunity to assist libraries in this regard by working collaboratively with them along with content providers, international standards organizations, our marketing partners, and local system vendors in meeting their users' demands for such information. We are certain library users, both now and in the future, will demand nothing less.

Library of Congress
December 19, 2000
Comments: lcweb@loc.gov

Librarians, Producers, and Vendors: The netLibrary Experience

by Lynn Silipigni Connaway, Ph.D.

Vice President of Research and Library Systems
netLibrary, Inc.

Final version

Introduction

In her presentation, "The Catalog as Portal to the Internet," during this conference on Wednesday, November 15, 2000, Sarah Thomas identified what's hot and what's not in terms of information and libraries. As I review this list, I can understand why vendors of electronic information and systems feel schizophrenic in today's environments. The web is hot, but libraries are not; eBooks are hot, but tree books are not; metadata is hot, but cataloging is not; portals are hot, but catalogs are not.

netLibrary, an eBook provider, offers published books on the web, but functions as a library in many ways, which I will discuss, and receives two tree book copies of every paper-published book that is offered as an eBook. Metadata is a term that is much used by netLibrary engineers, yet almost all of netLibrary's eBook metadata is provided by a machine-readable cataloging (MARC) record. netLibrary is considered a portal, but uses an Innovative Interfaces, Inc. cataloging system to track our eBooks and tree books.

What is an eBook?

There are many types of and definitions for eBooks. Some definitions, according to Walt Crawford [1], are: proprietary eBooks - Glassbook, Rocketbook; open eBooks - Open Ebook Forum specifications; public domain eBooks - Bartleby, Project Gutenberg; circulating eBooks - netLibrary; print-on-demand - Xerox, IBM, Sprout, Lightning Source, Hewlett Packard; vanity and self-publishing - various sources; diskette and CD-ROM - Modern Age Books; and extended books - Voyager.

What is a netLibrary eBook?

A netLibrary eBook most often has a print counterpart and has a defined beginning and end. It can be a monograph, reference book, edited volume, or multi-volume set. eBooks are searchable in two ways: within the specific eBook and across the collection of eBooks. The eBook can be enhanced with links and cross references to other electronic resources and with multimedia.

netLibrary provides an embedded look-up feature of Houghton-Mifflin's *The American Heritage(r) Dictionary of the English Language, Fourth Edition* that enables a user to highlight or double click a word and to view its definition and pronunciation while online. The Fourth Edition also provides pronunciations in audio format. An audio icon appears in the definition when a word pronunciation is available. When the audio icon is clicked, the user can listen to the pronunciation of the word. netLibrary provides digital rights management (DRM) software to protect its affiliated publishers' copyrighted content. This software allows users to copy and print limited information from within eBooks without violating copyright restrictions.

Why did netLibrary decide to catalog its eBooks?

netLibrary decided to provide cataloging for its eBooks because our library customers requested eBook MARC records. As a former head of a technical services department and a former cataloging professor, I did not fully understand or appreciate this request. I believed that cataloging multiple formats was the responsibility of professional catalog librarians, many of whom I had guided and educated to do exactly this.

As I continued to work with our library customers, I realized that it was imperative that netLibrary, as a vendor, partner with librarians to provide cataloging for eBooks. I was tasked with the responsibility of setting up a netLibrary technical services department.

The creation of the eBook MARC record allowed netLibrary to provide an eBook and MARC record package for our library customers. netLibrary was then able to develop relationships with bibliographic vendors, such as OCLC and RLIN, and library automation vendors, such as III, Sirsi, Epixtech, Endeavor, and DRA. These alliances enable our library customers and their users to access eBooks through their integrated library systems as they do with other library resources. The eBooks can not only be included in the online public access catalog module, but also in libraries' acquisition and circulation modules.

The inclusion of netLibrary eBook circulation statistics currently must be manually adapted to the circulation modules of the integrated library systems. This requires the library customer to dedicate time and resources to circulation integration. The integration of eBooks to acquisition and cataloging modules is more seamless.

Since netLibrary owns the eBook bibliographic records or metadata, it became apparent that this information could be used for internal processes. neLibrary uses the bibliographic data for: statistical analyses of collection and usage data; collection development that includes collection assessment and collection customization for library customers; and access points for the netLibrary search, retrieval, browse, and collocation.

As netLibrary began cataloging eBooks we soon could identify issues that have not been addressed by library bibliographic standards and formats, such as *Anglo-American Cataloguing Rules, 2nd ed. Rev. (AACR2R) and MARC Bibliographic Standards*. netLibrary's early work in eBook cataloging has given us the opportunity to assist in the development of eBook cataloging and metadata standards, such as our participation in this conference and our upcoming work with the newly formed OEB Metadata SIG and the AAP Metadata Standards ebook project. We have also recently been accepted as a NACO participant in the Program for Cooperative Cataloging (PCC).

What are the eBook cataloging challenges?

Cataloging Suppliers

The first challenge is to determine who should do the eBook cataloging - individual library catalogers, eBook providers, and/or bibliographic utilities. Today both individual library catalogers and eBook providers catalog eBooks and make them available through bibliographic utilities or directly through the eBook providers. Sharing the responsibility of cataloging enables all parties to be involved to efficiently provide cataloging for eBooks.

Standards and Schemes

The determination of metadata schemes and standards is currently one of the biggest challenges of cataloging eBooks. Publishers, librarians, technology providers, eBook distributors and vendors, and end users have some similar, yet distinct needs for describing and retrieving electronic resources. These distinct needs impact the schemes and standards used to code and display electronic resources. Librarians have traditionally used AACR2R to catalog bibliographic items and have encoded this information into the MARC format. Neither of these standards or formats has been widely used by publishers, technologists, or users. The Dublin Core was developed to identify metadata element sets for interoperability. ONIX was developed to identify and code the specifications utilized by the book trade industry. The OEB Forum has organized a Metadata SIG to review the most widely used standards, formats, and specifications.

As I continue to work with publishers and technology providers, it becomes obvious that there must be an integration of the different standards, formats, and specifications used to describe eBooks. This integration and collaboration is imperative to meet the needs of those associated with the creation, distribution, dissemination, and utilization of eBooks.

AACR2R has not adequately addressed the eBook format, although this may be the result of us, librarians, not interpreting the rules to accommodate eBooks. The *MARC Bibliographic Standards* seem to be more expansive and accommodating.

netLibrary Cataloging

At netLibrary all eBook cataloging is done with the print book in hand. The MARC record is part of the eBook package that is loaded to the netLibrary site. Since the netLibrary cataloging team catalogs only eBooks, its entire staff is trained and dedicated to this process. Both professional catalogers and cataloging assistants are trained and educated to copy catalog eBook titles included in the copyrighted and publicly accessible collections. Cataloging assistants are responsible for copy cataloging and professional catalogers are responsible for original cataloging.

The books format of the *MARC Bibliographic Standards* and chapter nine, computer files, of the *AACR2R* can be and are used for the cataloging of eBooks. The eBook is treated as a computer file and documented in the General Materials Designation (GMD). A 256 field also identifies the item as a computer file. Consequently, other cataloging conventions take precedence over the books MARC format. The GMD for the eBook will change from computer file to electronic resource, because of the work of the Committee on Cataloging: Description and Access (CC:DA).

Using these rules, there is no 300 field for physical description, although the netLibrary eBook retains the exact pagination and illustrative material as the print version of the book. A 007 field is used to describe the physical description of the eBook. A 538 note identifies the mode of access.

According to the cataloging rules, netLibrary is considered the publisher of the eBooks, although the content of all netLibrary eBooks is produced by a publishing house. netLibrary does not function as a traditional publisher, but converts the publishers' print or electronic files to the netLibrary electronic format. Regardless, netLibrary, Inc. becomes the publisher in the 260 field of the MARC record. Boulder, Colo. becomes the place of publication and the date of publication becomes the date of digitization.

A 776 note describes the publication information and physical description of the print copy of the eBook. Series notes are also included in this 776 field, since the 4xx fields are deleted in the e-book MARC record. The 4xx fields are deleted because the series statement identifies a print series, not an electronic series. A 500 note is included in the netLibrary MARC record to identify the original publisher and date of publication of the print book.

Publishers package many print books with supplemental materials, such as compact discs, maps, computer disks, realia, etc. It often is not possible to include these supplemental materials in the electronic environment of the eBook for various reasons, which may include rights, which brings up another inadequacy of *AACR2R* in cataloging eBooks. There is no provision for digital rights management (DRM) information. Catalogers must attach disclaimer notes to the netLibrary MARC records for eBooks that do not contain the supplemental information that is available with the print version of the book.

The use of chapter nine of the AACR2R may not concern libraries that follow the *Library of Congress Rule Interpretations (LCRI)*. OCLC, on behalf of netLibrary, worked with the Library of Congress and proposed LCRI 1.11A, which treats an eBook as a reproduction of the print version, much as a microform is treated in relation to the print version of a work. The LCRI allows the 260 field to retain the print publisher information, the physical description of the book to be retained in the 300 field, and the series statement to be retained in the 4xx fields. A 533 field, reproduction note, is added to identify the item as an electronic reproduction and to document the electronic publisher, date of digitization, and the mode of access.

Holdings Records

A library can attach its holdings to a record for netLibrary eBooks because it purchases them just as the library purchases print books. If, when attaching holdings to an eBook, a location is required for display in the integrated library system, the determination of the location should explicitly identify the item as an electronic full-text book available on the Internet through a browser. Attaching holdings to eBooks may further confuse staff and patrons when requesting materials through interlibrary loan (ILL) systems. If the eBook contract does not allow for ILL, this also must be clearly documented in the record.

Collocation

Linking bibliographic items can also be challenging in the eBook environment since it is optimal to link the text about an item with photographs, images, and audio and video segments of the item. Determining and including identifiers in the MARC record, such as the Uniform Resource Locator (URL), Visual Resources Association Core (VRA), Persistent Uniform Resource Locator (PURL), Digital Object Identifier (DOI), and Encoded Archival Descriptions (EAD), will greatly enhance the collocation of bibliographic items.

Incorporating authority control, controlled vocabularies, and subject codes, such as BASIC/BIC and the *Library of Congress Subject Headings (LCSH)*, will allow users from different backgrounds and disciplines to retrieve more relevant and precise information. Applying classification numbers to eBooks also enables collocation and precision, although it is time consuming and may cause confusion for the users. There is a concern among librarians that if classification numbers for eBooks are displayed to users, they will go to the physical shelves of the library to secure the items and be discouraged and frustrated when these items are not on the shelves.

Statistics

Unanswered questions still remain concerning the reporting and counting of eBooks. The Association of Research Libraries (ARL) currently is developing policies for libraries to count and report eBook statistics. Since libraries can purchase netLibrary eBooks in perpetuity, experts believe that these eBooks should be counted and reported as expenditures and volumes added to the collection and as part of the circulation statistics, as other formats and types of materials are counted and reported. Subscription eBooks may be counted and reported as expenditures, but not as volumes added to the collection.

Staffing and Processes

Technical services department librarians need to evaluate their current processes and staffing before integrating eBook cataloging into their work flows. A library can merge eBook cataloging into its centralized cataloging processes, determine whether the titles are monographs or serials, and process them accordingly. A specialized electronic resource librarian or a media cataloger can handle acquisitions, licensing, and cataloging of eBooks. The technical services manager must decide whether both copy and original catalogers should catalog eBooks, or if one cataloger is sufficient to handle both copy and original cataloging. Staffing and workflow must be tested to identify the best solution for each technical services department.

Training and Education

Library staff, as well as library patrons, must be educated and trained to access and utilize eBooks. Librarians often are most concerned with educating and training patrons; they forget about the education and training of staff. Staff will accept new responsibilities, new formats, and new technologies more readily if they are comfortable with and knowledgeable about the changes. If staff are confident handling new technologies, they become more apt to share their enthusiasm and expertise with library patrons.

What are netLibrary eBook cataloging challenges?

As stated above, the netLibrary cataloging team has experienced changes within the past eighteen months. The cataloging team physically moved to the netLibrary Publishing Building with the production team. The cataloging team works closely with production engineers to create an eBook and MARC record package for our library customers and for internal use.

We have become very involved and interested in the assignment of ISBNs to the different formats of eBooks, such as netLibrary, Peanut Press, Glassbook, Rocketbook, MetaText, Microsoft, etc. We work with several representations of different electronic versions of the same material. This goes back to an issue discussed throughout this conference - the content vs. the carrier.

netLibrary has decided to create and store separate and distinct records for each format, because we may need them separated in the future. If a library wishes to combine records for different formats, it is done at the institutional level. Thus, netLibrary may have multiple records for any given title.

netLibrary has created eBook collection sets, such as the *Choice* Outstanding Academic Titles and the Oxford University Press collections. Ideally, these collections can be identified in some way within the MARC record to collocate all titles within these collections, without jeopardizing the integrity of the individual eBook titles' MARC records for those libraries that do not purchase the entire collection sets. This is only one example of the challenges faced by a vendor that is cataloging for multiple libraries and multiple systems, with different requirements and needs.

netLibrary delivers eBook MARC records to multiple library customers who use various integrated library systems. These systems have various indexing and loading requirements so netLibrary must work closely with the integrated library system vendors and our library customers to facilitate the loading of eBook MARC records. This means that netLibrary staff must be knowledgeable in the requirements of the various integrated library systems and maintain accurate and up-to-date help files and FAQs for library customers.

Quality control requires time, as well as human and technology resources. As standards for eBook cataloging change, netLibrary must be prepared to change our cataloging practices and processes, as well as retroactively change all MARC records and make them available to our library customers.

What are netLibrary eBook cataloging benefits?

With all of the challenges associated with eBook cataloging mentioned above, one may ask why netLibrary or any other eBook provider chooses to embark on this endeavor. netLibrary began cataloging eBooks in response to our library customers' requests. We believe it is imperative for librarians, publishers, bibliographic utilities, vendors, and eBook providers to work together to integrate eBooks into the digital library. Cataloging eBooks has not only enabled netLibrary to work cooperatively with for librarians, publishers, bibliographic utilities, vendors, and other eBook providers, but has also provided us with metadata that streamlines our internal processes and information retrieval on our site.

The netLibrary cataloging team has cataloged approximately 39,000 eBooks since April 2000. We have delivered over 75,000 eBook MARC records to libraries, vendors, and bibliographic utilities. These figures alone indicate that eBook cataloging does demand partnerships and cooperation. No one entity could possibly accomplish this alone.

What's next?

I believe that we must utilize the capabilities of the eBook. It is more than an alternative to a paper book. We, librarians, must think beyond the paper book. Let us not make the mistake that we made when moving the paper card catalog to the online environment - simply digitizing the catalog card, without considering the new possibilities for search and retrieval. We should include links from the eBook to dictionaries, thesauri, related images, photographs, electronic text, and audio and video segments.

Now is also the time to enhance the bibliographic record. We should utilize the table of contents and book indices in the bibliographic record since these are already digitized in the eBook format. We should also include links to book reviews, electronic resources that are referenced in the book, and book summaries. We need to work with publishers, technology providers, and eBook providers to not only map standards and schemes, such as the Dublin Core and ONIX, but to integrate these into the MARC format.

The incorporation of full-text search capabilities of eBooks should be integrated into our library online public access catalogs to enable users to search within the library's electronic collection, as well as across other available electronic collections. CORC can be used as an example to move in this direction, since it enables users to search across all types of electronic information, i.e., web sites, electronic journals, eBooks, newspapers, advertisements, etc. Library systems should also enable the integration of semantic searches that map and retrieve concepts and ideas in addition to keyword and known searches.

These advances will move libraries into the digital world of our users. With the advancement of wireless technologies available through Yahoo, AOL.com, and car manufacturers, library users' expectations are changing and they are more wired and more dependent upon technology. E-cars, high-tech automobiles with Internet access, will allow individuals to check e-mail, monitor stocks, and keep up with sports scores without taking their hands off of the steering wheel because of telematics, a new wireless technology that transmits information to and from a vehicle. Telematics is available in 2001 automobiles from Mercedes-Benz and General Motors and includes voice-activated features.[2]

The popularity of napster and MP3 have given users the capability to aggregate their electronic content into private digital libraries. The popularity of peer-to-peer technology, such as gnutella, fashioned after napster but that allows all types of files to be shared between individuals, is facilitating this aggregation.

If individuals are aggregating content to create their own information stores, will libraries and librarians become obsolete? The literature indicates that librarians will be needed to assist individual users with the retrieval and evaluation of electronic information.[3] John Lombardi also anticipates that the role of the librarian as gatekeeper will change as individuals become their own gatekeepers. He believes that librarians will digitize unique special collections and maintain and manage these collections. He also envisions librarians creating a "mega" library union catalog and developing library portals to compete against commercial services.[4]

In her presentation at the Computers in Libraries 2000 Conference, Rebecca Jones of Dysart and Associates, stated that librarians will not be "disintermediated by end-users searching the Web since search Web search engines index only 55% of the web." Rebecca believes that librarians will function as "metamediaries."[5]

John Lombardi has outlined his "Rules for Digital Survival." They are: objects are not as important as the content; helping clients find resources in a "digitally chaotic world is the first priority"; and "for the next ten years, if it works well, is reliable, and you know how to use it, it is obsolete"[6] With this, I would like to end with a quote from *Future Shock*, because I believe that if we, as librarians, adhere to this quote by Toffler, we will become obsolete. "The illiterate of the year 2000 is not the one who cannot read and write, but the one who cannot learn, unlearn and relearn."[7]

---

1. Crawford, Walt. "Nine Models, One Name: Untangling the eBook Muddle." *American Libraries* (September 2000): 56-9.
2. Hales, Dianne. "E-Cars take to the Road." *Parade Magazine* (October 1, 2000): 18-9.
3. Keller, Larry. "Looking It Up." (November 28, 2000). http://www.cnn.com/2000/CAREER/trends/11/28/librarians/index.html.
4. Lombardi, John. "20/20 Vision for the Future." Paper presented annual meeting of the American Library Association, University Libraries Section and The College Libraries Section of ACRL. Chicago, July 2000.
5. Jones, Rebecca. "The Library of the Future and the World Network." Paper presented at the

Computers in Libraries 2000 Conference. Washington, D.C., March 15-17, 2000.
6. Lombardi, John. "20/20 Vision for the Future."
7. Toffler, Alvin. *Future Shock*. New York: Random House, 1970.

Library of Congress
December 19, 2000
Comments: lcweb@loc.gov

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Regina R. Reynolds

Head, National Serials Data Program
Library of Congress
101 Independence Ave., S.E.
Washington, DC 20540

Partnerships to Mine Unexploited Sources of Metadata

About the presenter:

Regina Romano Reynolds has been head of the National Serials Data Program, the U.S. ISSN center, since 1992. Reynolds has worked at the Library of Congress since 1976 and has spent much of her professional career explaining and promoting ISSN use to publishers and the information community. Reynolds has an M.L.S. (Beta Phi Mu) from the University of Michigan. She is active in the American Library Association and the North American Serials Interest Group where she is a frequent author and presenter on topics in serials, standards, and electronic resources. Reynolds was the 1999 recipient of the Bowker/Ulrich's Serials Librarian Award. She is also actively involved in the revision of AACR2 to accommodate seriality and electronic resources as well as in the international harmonization of cataloging rules and standards.

Full text of paper is available

Summary:

If the library catalog is to play any role as a portal to Web resources, new means have to be developed to bring the ever-increasing number of Web resources of interest to library patrons under some kind of bibliographic control. Traditional cataloging of published textual materials, which has

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

been largely monolithic to date, will have to be progressively subdivided into a hierarchy of different record levels aligned with the research (and probably monetary) value of the resource. At the highest level, traditional cataloging will still prevail. At the lowest level, records might be produced from publisher-supplied or secondary-source metadata which has been formatted into MARC records for inclusion in library catalogs. These metadata-based records could also be selectively edited by trained catalogers and optionally enhanced with authoritative name and subject headings. OCLC's CORC program is one opening wedge to the entry of such non-AACR-based records into shared databases and library catalogs.

To realize fully the potential of such metadata-based catalog records, new partnerships and new sources of cataloging data have to be explored and exploited. Metadata created in association with existing identifiers such as the ISBN and ISSN, and metadata planned to support emerging identifiers such as the Digital Object Identifier (DOI) and the developing identifier, the ISTC (International Standard Text Code), are potential sources of bibliographic data which libraries can convert, or convert and enhance to produce MARC records. Non-identifier-based publisher registration procedures such as CIP, Copyright, and others might also yield useful data. As all of these registration procedures are increasingly completed electronically, they yield data which are highly manipulable, enhanceable and convertible.

In addition to exploring sources of metadata, especially metadata supplied by publishers as part of registration procedures, this paper will examine ways in which such registration procedures could be modified to better provide libraries with needed data. Such modifications include addition of elements needed for library cataloging, and provision of instructions which will result in publishers providing data in more standardized ways. With the increasing use of online forms, interactive programs could be developed to "talk" publishers through the process of completing registration forms in such a way as to make them more useable for conversion to basic catalog records. Finally, ways for publishers to provide subject information will be explored.

The potential for creation of catalog records based on publisher-supplied metadata will be illustrated using data from a study of records created by the National Serials Data Program (NSDP). NSDP, the U.S. ISSN Center, uses an online form for ISSN registration. Publishers complete the form according to instructions supplied by NSDP. Data from the online form is converted to a draft catalog record which is then edited and enhanced by professional catalogers. Results of a study of the usability of information supplied by publishers on the ISSN application form, and the editing required on NSDP records produced by conversion from the online application form, will be presented.

Library of Congress
May 9, 2000
Comments: lcweb@loc.gov

Partnerships to Mine Unexploited Sources of Metadata

Regina Romano Reynolds

Head, National Serials Data Program
Library of Congress

Final version December 2000

The old order changeth, yielding place to new,
And God fulfils himself in many ways,
Lest one good custom should corrupt the world.
"The Passing of Arthur," Alfred Tennyson

"The high cost of traditional library cataloging makes it impractical in the context of such growth and less expensive alternatives are needed for many, if not all of those resources..." [1]
"Recommendation: The Library should actively encourage and participate in efforts to develop tools for automatically creating metadata. These tools should be integrated in the cataloging workflow." [2]
LC 21: A Digital Strategy for the Library of Congress

Introduction

Let me not be quoted as saying that the "good custom" of cataloging as refined and practiced for over 100 years by the Library of Congress has "corrupted the world!" Nevertheless, as many of the conclusions of the National Academy of Science report, *LC 21*, indicate, it is becoming ever more clear that the old order of cataloging must, indeed, yield place to a new one, no doubt a multi-faceted one, with many levels, schemes, players, and partners.

If the resources that are being put onto the Web each day were, instead, being published in print form and being received into the Library of Congress, there would not be enough room left in any of our three buildings for us to hold this conference! Even though these backlogs are intangible, they nonetheless continue to grow to the point where we must think of new solutions--solutions which this conference is being held to explore.

Even though I feel we need new solutions, I do not feel that cataloging is either unnecessary or obsolete. At least not in the near-to-medium term. There may come a day when information is self-indexing, when discovery mechanisms will have progressed to the point where there is no need for this traditional library function, but, as the papers prepared for this conference indicate, we are not there yet. The thoughts I offer here are for the short-to medium term, in order to transition from where we are to where we might eventually go.

We have created our library catalogs over the course of hundreds of years, and these catalogs provide access to many resources which are not yet in digital form and some which might never be digitized. As Sarah Thomas and others have already indicated, in order to provide access to the collective wisdom the library contains we have to find ways to integrate the discovery of both traditional and Web-based resources into our catalogs. One of the gatekeepers to this process is the restriction of many catalogs to AACR and MARC-based records. CORC, with its ability to translate Dublin Core elements into MARC tags, is providing an opening wedge for non-AACR-based records.

One potential solution which I and various others have been advocating for many years is to widen the concept of levels of cataloging to encompass a hierarchy of different record levels aligned with the research (and probably monetary) value of the resource. At the highest level-rare books and materials of high research value, for example--traditional cataloging by experts in subject and descriptive cataloging would still prevail. At the lowest level-Web-based resources of moderate research value which might not otherwise receive any bibliographic control--records might be produced from publisher-supplied or secondary-source metadata which has been formatted into MARC records for inclusion in library catalogs. At levels in between these two extremes, a combination of automated and cataloger-supplied data would be used. For example, metadata-based records might be augmented by authoritative name and subject headings. Or, metadata-based records might be edited by catalogers or cataloging technicians as well as being enhanced with authoritative name and subject headings.

How can libraries obtain the metadata they need to implement such a hierarchy of cataloging levels? When I thought about this question, I realized that obtaining metadata is a lot like getting money: you can inherit it, get it the "old-fashioned way" by earning, or creating it, or-marry it!-that is, find a partner with a lot of it. And, what is the modern way to find a partner? Put an ad in the personals! So, I took the liberty of writing the following ad for the Library of Congress:

```
        200-year-old Library (looks 100), mature, experienced, with millions of
assets
        seeks young, exciting, digitally-savvy partners with ample metadata to share.
        We can make beautiful catalogs together!  Willingness to convert to MARC a
plus.
```

Although this example is obviously facetious, how libraries can obtain the metadata upon which to build a hierarchy of different cataloging levels is the subject of this paper. My premise is that metadata created for other purposes--particularly metadata created in association with existing and emerging identifiers--ISSN, ISBN, DOI--or captured from registrations such as Copyright or Cataloging in Publication, are potential sources of bibliographic data, and that non-traditional solutions such as this are the only way libraries will be able gain some bibliographic control over the explosion of Web-based resources. Partnerships with agencies which collect this metadata can provide opportunities to share libraries' experience using metadata so as to make it readily adaptable for library cataloging purposes.

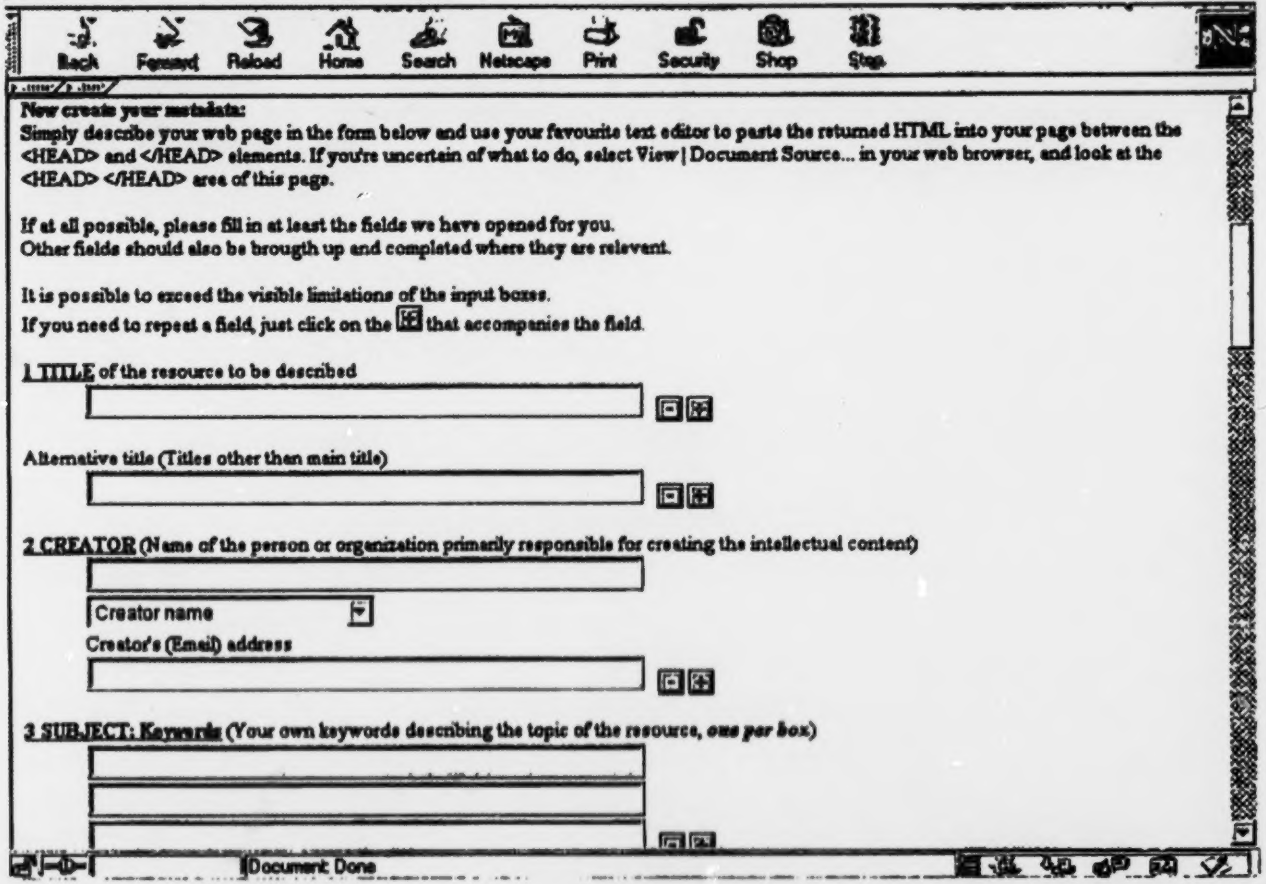
Collecting Metadata by Means of Templates

"The creation of more and better metadata--structured resource descriptions either embedded into documents themselves or external to them--is generally regarded as the best means to improve the current situation. Many specialists believe that any metadata is better than no metadata at all--we do not need to stick with the stringent quality requirements and complex formats of library catalogue systems. Instead it is possible to live with something simple, which will be easily understandable to publishers, authors, and other people involved with the publishing of electronic documents." The Nordic Metadata project. Final Report, Introduction. 3.

If all producers of Web-based resources created metadata according to a rich and accepted standard and embedded this metadata into the HTML header of their document, the "automatic" creation of metadata for library catalogs recommended by *LC 21* would be well on its way to becoming a reality. If XML and RDF were in widespread use and implemented by Web browsers, we would also be closer to "automatic" metadata creation. Again, we are not there yet. Simply reacting to the potential of existing technology will not yet get us where we want to go. Therefore, libraries must become more proactive in pursuit of the means to control Web resources.

One approach in this direction has been the use of templates (online fill-in forms) to collect metadata elements and the development of programs to format the collected metadata into standardized syntax, such as Dublin Core HTML. This is the approach taken by, most notably, the Nordic Metadata Project and BIBLINK. Briefly, the Nordic Metadata project, which ran from 1996 - 1998 chose the Dublin Core element set as its metadata format, developed a template to collect and format metadata, developed metadata harvesting and indexing applications, a Dublin Core-to-MARC converter, and a URN generator. BIBLINK, a European Commission-sponsored project which ran from April 1996 - February 2000 had as its aim "to establish a relationship between national bibliographic agencies and publishers of electronic material, in order to establish authoritative bibliographic information that would benefit both sectors." 4. BIBLINK used a template as a mechanism by which publishers could send metadata to national bibliographic agencies.

Nordic Metadata Project's Dublin Core Metadata Template



Now create your metadata:
Simply describe your web page in the form below and use your favourite text editor to paste the returned HTML into your page between the <HEAD> and </HEAD> elements. If you're uncertain of what to do, select View | Document Source... in your web browser, and look at the <HEAD> </HEAD> area of this page.

If at all possible, please fill in at least the fields we have opened for you.
Other fields should also be brougth up and completed where they are relevent.

It is possible to exceed the visible limitations of the input boxes.
If you need to repeat a field, just click on the ⊞ that accompanies the field.

1 TITLE of the resource to be described

Alternative title (Titles other then main title)

2 CREATOR (Name of the person or organization primarily responsible for creating the intellectual content)

Creator name

Creator's (Email) address

3 SUBJECT: Keywords (Your own keywords describing the topic of the resource, *one per box*)

Use of templates to collect and standardize metadata from resource creators, as in the BIBLINK project above, can result in two distinct outcomes, both highly relevant to library bibliographic control. First, by use of "crosswalks," such as those developed to convert Dublin core elements to MARC elements, MARC records can be output from data input on such templates and integrated into OPACS. Second, Dublin Core metadata coded according to HTML standards can be output and returned to the resource originator to be included in the HTML head of the resource, thus enabling cataloging tools such as OCLC's CORC and others to produce much more complete and immediately-usable MARC records.

Although the potential exists to extract some metadata from resources themselves, especially those where metadata has been embedded by the resource creators, one premise of this paper is that for the short-term, information elicited by carefully designed templates will be much more compatible with existing catalogs and the standards by which most of those catalogs have been created. In the report, "Project BIBLINK: Linking Publishers and National Bibliographic Agencies," Manjula Patel and Robina Clayphan state, "Records produced using the Web Interface have been of a better standard than those extracted from data already embedded in some on-line resources. The Web form has the advantage from the library's point of view, of concentrating attention on the task, limiting errors that can be made, and causing the user to create the record *following the guidelines provided.*" 5.

BEST COPY AVAILABLE

Potential Partnerships for Re-purposing Metadata for Bibliographic Use

Potential partners exist both within and outside of the library community. The Library of Congress is in a unique position to build on the experience of projects such as the Nordic Metadata Project and BIBLINK. The Library already receives metadata--some of it in digital form--in conjunction with at least three registration processes under its control. The U.S. Copyright Office, the Cataloging in Publication program, and the National Serials Data Program (NSDP, the U.S. ISSN center), all use registration forms on which publishers supply metadata. All three programs have mechanisms to accept at least a portion of their registrations using online forms. Outside of the Library, emerging projects with potential include the DOI for metadata created in connection with the Digital Object Identifier; registration agencies to be created in conjunction with the proposed ISO standard, the ISTC (International Standard Text Code); the ISBN Core Metadata project; and metadata collected to support OCLC's Open Name Services project. These emerging projects are very much still under development, so their potential is difficult to predict. On the other hand, opportunities for collaboration and helping to shape the metadata which is collected might be greater with emerging projects than with the well-established programs the Library of Congress is associated with.

U.S. Copyright Office

The *LC 21* report noted, "The Library's role in registering copyright and enforcing the mandatory deposit law creates a unique opportunity for it to collect digital information that might otherwise vanish from the historical record." 6. Along with this unique opportunity to collect digital material, the Library also has a corresponding unique opportunity to collect metadata to support not only the copyright registration but also to facilitate the bibliographic control of this material. Although at present most copyright registrations are made by applicants using paper forms, an "Electronic Registration, Recordation & Deposit System," CORDS, is under development and has been in limited use with a group of publishers who use a digitized template of the paper form to submit applications and copies of digital materials.

As CORDS is moved into a production system, as urged by the LC21 report, or as a new digital registration system is developed (another option presented by *LC 21*), the potential for collaboration with the cataloging operation of the Library will open up in a new way, although some significant hurdles will need to be overcome. At present, cataloging of copyright registrations and cataloging for bibliographic control have been separate, parallel systems at the Library of Congress. For years, those not familiar with the details of copyright registration have wondered why these operations could not somehow build upon each other. At first glance, this seems like an obvious possibility. However, Associate Register Mary Levering explained to this author that the copyright form collects only "copyright registration facts," metadata which do not necessarily match metadata created for traditional cataloging. 7.

A preliminary comparison of the Dublin Core elements with the elements collected as part of the copyright registration process reveals the following:

D.C. Element	**On Copyright Form**
Title	Title
Creator	Author
Subject	----
Description	----
Publisher	----
Contributor	Combined with author

Date	Date creation of work completed Date of first publication
Type	Type is denoted by choice of form, e.g., VA= visual works, PA = performing arts works, etc.
Format	Some formats require specific forms
Identifier	URL is not collected; ISSN is collected only on form SE/Group which is used only for group registration of periodicals
Source	Derivative work or compilation
Language	----
Relation	Title of collective work for works published as contributions to periodicals, serials, or collections, includes designation and pages of the periodical or serial
Coverage	----
Rights	Claimants, OPTIONALLY on some forms: name and address of person to contact re rights and permissions

Additional elements present on Copyright Registration Forms

Author's nationality or domicile
Whether the author's contribution was anonymous or pseudonymous
Author's dates of birth and death
Nature of authorship (entire text, co-authorship, compilation, translation, etc.)
Address of Copyright claimant
Previous registration
Work made for hire

The above brief comparison indicates that for five of the 15 DC elements there is no equivalent data and for various others the data is incomplete or does not map completely. Such typical bibliographic elements as subject, place of publication and publisher are totally absent. Nonetheless, because much work remains in order for the Copyright Office to have a production system for the registration of electronic materials, potential for collaboration remains. Additionally, both the Library's cataloging operations and Copyright Office have a common interest making the best use of limited staff in the face of the increasing volume of electronic materials confronting each operation. A form of collaboration heretofore seen as only desirable might soon become essential.

Associate Register Levering indicated that the Copyright Office could ask for additional elements on the registration application--especially if provision of such elements was optional-- without any change in law or regulation. However, she also indicated concern about placing a potentially discouraging burden of information on Copyright applicants. A first step would be providing the opportunity for publishers to supply additional optional metadata which would support the Library's cataloging operation. Currently, certain copyright forms provide a box (labeled OPTIONAL) in which to provide "Name and Address of Person to Contact for Rights and Permissions." A very informal estimate by some Copyright Office staff indicates that this optional information is often provided. Levering indicated that consensus support for the inclusion of provisions for the collection of additional metadata-probably optionally-- from key organizations in the library, publishing, and information communities could lead to consideration of including these additional elements in a future digital copyright registration system. 8.

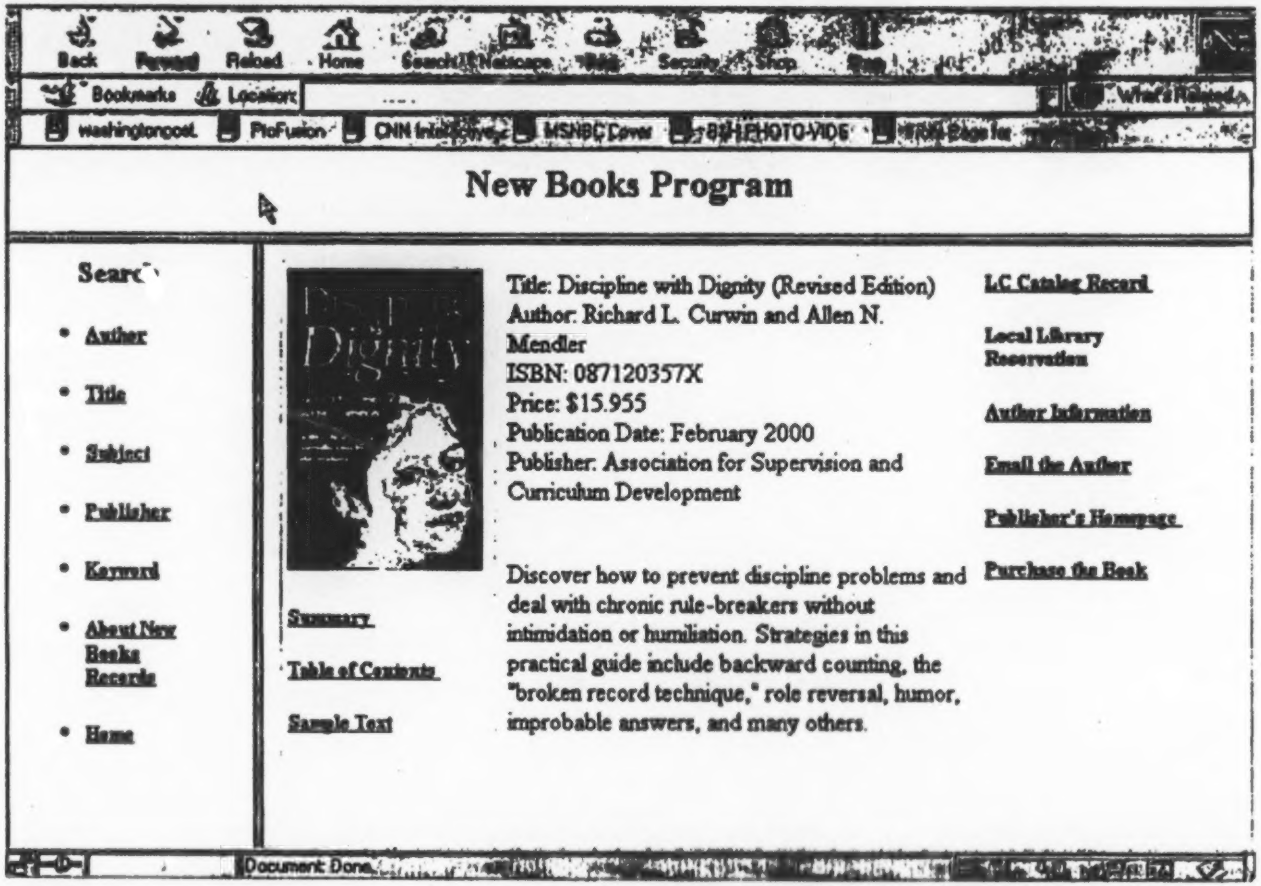
Once the minimum elements needed for a library catalog record were collected via an electronic copyright registration system template, these data could be converted into baseline records and shared with the bibliographic community as LC has been doing with other kinds of cataloging data for the over a century. The challenges to be overcome are considerable but the payoffs in terms of increased control of U.S. digital resources would be great.

Cataloging in Publication Office (CIP)

Another of the "narrow gates" through which many U.S. monographic materials pass, and thus a point for the potential capture of metadata for electronic materials, is that of CIP. The current CIP registration form elicits much more bibliographic information than the Copyright form. However, at present, Web resources are not included in the CIP program. The "Electronic CIP" program, thought of by some as pertaining to digital materials, in fact, only encompasses printed materials sent to LC in digital form but, given publishing and library control trends, that policy would seem to be destined for change in the future. Nonetheless the present "Electronic CIP" program does provide a model for the receipt of digital materials in the future since the Electronic CIP programs designed to process applications for printed books could easily be adapted for use with digital resources. Additionally, a proposal entitled "New Books" currently under development in the CIP office also demonstrates the power of the CIP program to interact with publishers for a net result with great potential for the control of electronic resources.

Under the Electronic CIP Program, publishers complete an online form and attach a digital file of the book. Using a program developed at the Library, the data on the form is converted to a draft MARC record. LC catalogers use the draft record and digital file to produce cataloging in publication records, which are later updated using donated copies of the published books. Working with galleys in digital rather than printed form has produced some distinct advantages. More CIP records are based on full text rather than simply front matter; fewer typographical errors occur because catalogers can cut and paste data from the electronic text; and quicker throughput time results from eliminating transit time in the mail and within the Library.

The proposed New Books program is still just a gleam in Cataloging in Publication Program Chief John Celli's eye--and a prototype on an internal Library server. Nonetheless, the prototype dramatically demonstrates the usability of publisher-supplied metadata to create not only basic catalog records but also to enhance such records in ways which the public is coming to expect based on their use of such Web sites such as Amazon.com.



New Books Program

Search

- Author
- Title
- Subject
- Publisher
- Keyword
- About New Books Records
- Home

Title: Discipline with Dignity (Revised Edition)
Author: Richard L. Curwin and Allen N. Mendler
ISBN: 087120357X
Price: $15.955
Publication Date: February 2000
Publisher: Association for Supervision and Curriculum Development

Discover how to prevent discipline problems and deal with chronic rule-breakers without intimidation or humiliation. Strategies in this practical guide include backward counting, the "broken record technique," role reversal, humor, improbable answers, and many others.

Summary

Table of Contents

Sample Text

LC Catalog Record

Local Library Reservation

Author Information

Email the Author

Publisher's Homepage

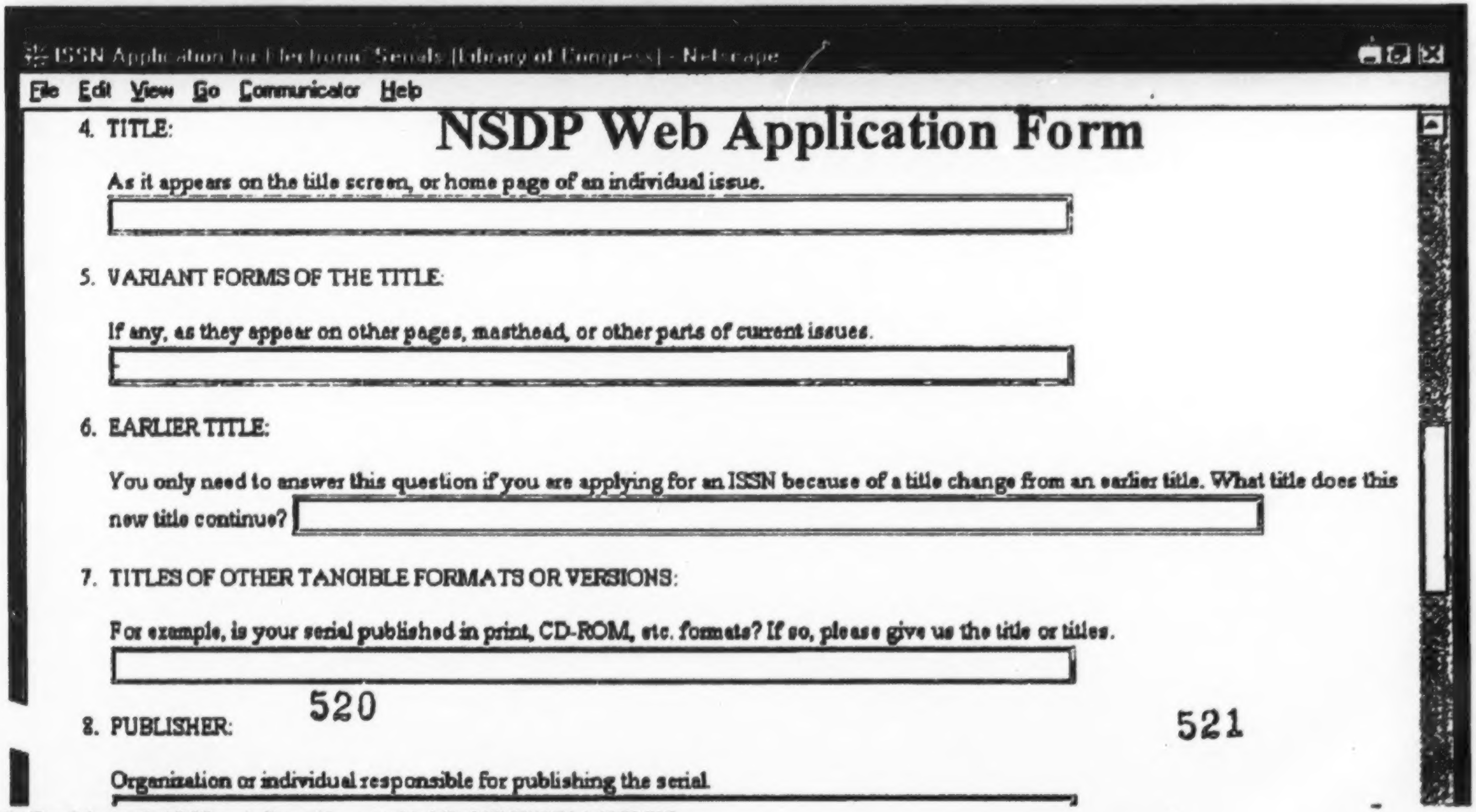
Purchase the Book

The New Books proposal would provide for a template on the Web which participants in both the CIP Program and the Preassigned Card Number program could complete. Elements requested on the template are the traditional catalog record elements such as title, subtitle, place, publisher, date, subject, audience, ISBN, plus the potential to include a summary, reviews, table of contents, and image of the book jacket. The data input on the template and the attached files are then converted into a record that looks like a cross between a traditional catalog record and a listing on Amazon.com. Records created in this manner would, under the New Books proposal, form their own adjunct to the LC OPAC as well as reside in a space reserved for them on LC's Web page. As the materials were published and selected for inclusion in LC's collection, they would receive traditional cataloging and records would be removed from the Web site under one scenario, or simply moved to a published portion of the site under another. Although not intended, as now conceived, as a replacement for traditional cataloging, it is difficult to imagine that the potential efficiency to be realized by simply bringing this publisher-created record under name and subject control might not lead to such a future. Furthermore, although the initial New Books concept does not focus on digital materials, given the pressures on the Library of Congress brought to bear by the challenge of collecting and controlling a much greater percentage of Web-based resources in the future, it is also hard to imagine that the potential for this kind of record-creation mechanism will not be recognized and extended to digital resources. John Celli, himself, has indicated to this author that digital resources will probably have to be included if the project is to gain widespread Library support.

One very attractive feature of the New Books prototype is the potential for the addition of subject headings based on a pull-down menu of the "BASIC Subject Heading Codes," a list developed by the Book Industry Study Group which will be discussed at greater length in the last section of this paper. This list consists of "over 2500" subject headings--a tiny fraction of the number in the 5-volume set of Library of Congress Subject Headings (LCSH). Nonetheless, publishers' use of this list would provide controlled subject vocabulary, and either alone, or better yet, coupled with key word searching, would result in some controlled vocabulary subject access without intervention by library cataloging staff.

ISSN Registrations

Since 1996 the National Serials Data Program (NSDP), the U.S. ISSN center, has used a template form on its Web site (http://lcweb.loc.gov/issn) for ISSN registration of U.S. digital serials. The online form includes the same data elements as those present on the printed application form but its use is restricted to digital serials. Use of the online form requires the publisher to provide either digital files along with the application or a URL so that ISSN center staff can view the serial in order to determine its eligibility for registration. The online form already provides many advantages to quicker and more accurate ISSN registrations and records for digital serials but it holds even more potential for the future, since both the template and conversion program are still quite basic and only the ISSN is returned to the publisher at present.


ISSN Application for Electronic Serials (Library of Congress) - Netscape
File Edit View Go Communicator Help

NSDP Web Application Form

4. TITLE:

As it appears on the title screen, or home page of an individual issue.

5. VARIANT FORMS OF THE TITLE:

If any, as they appear on other pages, masthead, or other parts of current issues.

6. EARLIER TITLE:

You only need to answer this question if you are applying for an ISSN because of a title change from an earlier title. What title does this new title continue?

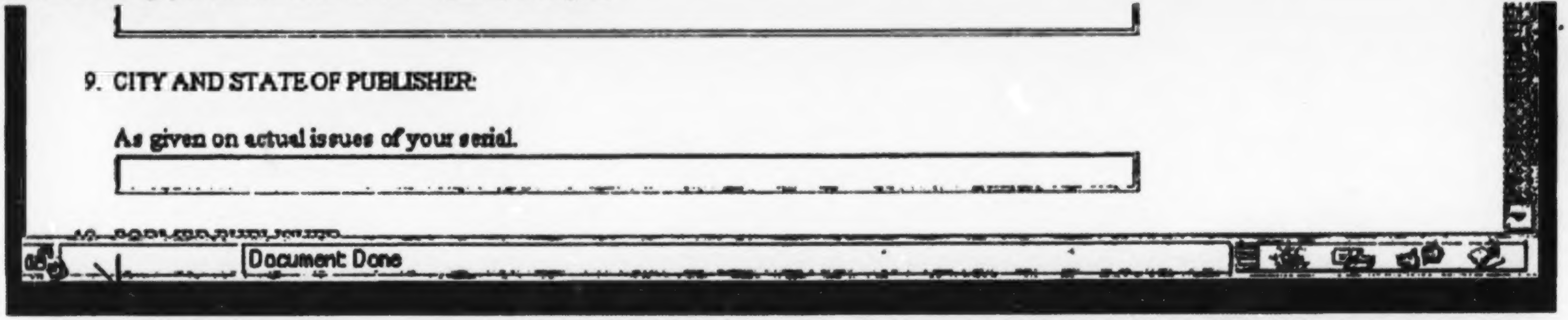
7. TITLES OF OTHER TANGIBLE FORMATS OR VERSIONS:

For example, is your serial published in print, CD-ROM, etc. formats? If so, please give us the title or titles.

8. PUBLISHER:

Organization or individual responsible for publishing the serial

BEST COPY AVAILABLE



9. CITY AND STATE OF PUBLISHER:

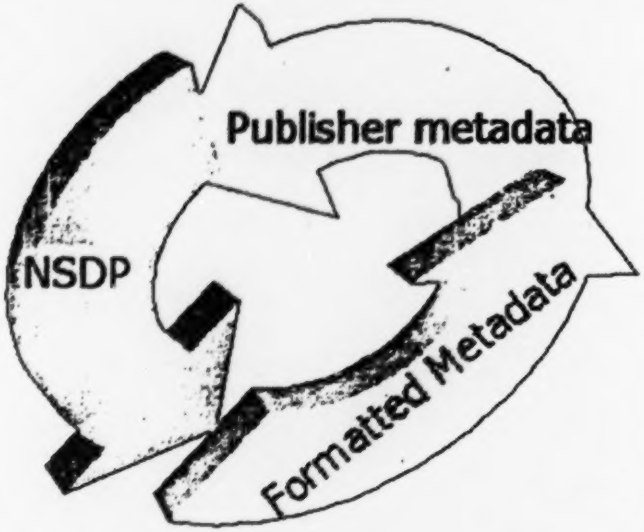
As given on actual issues of your serial.

P functions within the CONSER program, a cooperative cataloging program for U.S. serials. As part of CONSER, NSDP creates CONSER records in the OCLC database for the serials to which it assigns ISSN. The conversion program has been designed for use with OCLC. To begin with, the template program screens applicants by asking if the resource for which registration is an online serial, and whether it is published in the United States. If either of these questions is answered in the negative, an error message appears, and the form cannot be submitted. Instead, instructions are provided for how to register serials in other formats or how to contact other ISSN centers to register serials published outside of the United States. Thus, inappropriate registrations are avoided--for the most part. Of course, there is always a way for a determined applicant to subvert even the most well-designed program. For example, NSDP recently received, via the U.S. Mail, a printout of NSDP's online application form completed by a library organization in the United Kingdom. The form was accompanied by a note which indicated that the organization could not successfully submit the form online and thus were sending it by post!

When the cataloger processes the file resulting with the data supplied by the publisher, the elements on the form which contain data appropriate for a MARC record are mapped to their corresponding MARC21 fields by a program developed at LC. The cataloger then edits and augments the resulting draft record according to the requirements of the ISSN Network, AACR2 and MARC21, including the checking of authority files. The current program is quite basic--more elements could be requested from the publisher, more explicit instructions could be given to obtain better data, and better error checking could be done. Nonetheless, catalogers find the program cuts down on the keying they have to do, and, even though publishers can also mis-key data, use of the program has resulted in greater accuracy, especially in URL fields. Additionally, processing efficiencies have been realized since there is less paper to move, file, and potentially misplace; and publisher notification of the ISSN is done by an email program, thus reducing notification time and effort.

The potential for returning standardized and formatted metadata to resource creators is an exciting byproduct of the registration process. NSDP and other registration agencies are in a unique position to complete the information loop by returning metadata back to the publisher--now standardized, enhanced, and properly formatted for embedding into the head of the resource. NSDP is hoping to develop a program--or to use OCLC's CORC capabilities--to output not only a MARC21 record for cataloging but also HTML metadata for return to publishers. Publishers who are standards-aware enough to have requested ISSN might be especially good candidates for embedding standardized metadata into their resources if it is supplied to them at the time of ISSN registration. In this way, search engines, harvesters and programs such as MARC-it and CORC can return much superior results.

Linking Publishers & Libraries



NSDP receives approximately 50 ISSN applications a month using the online template. This volume of template use has taken place with no specific publicity on the part of NSDP. The number of online registrations would probably be much greater if this capability were publicized but NSDP lacks the staff to keep up with even the current number of ISSN requests. The ISSN Network is also exploring the potential of template-based self-registration by creators and publishers of digital resources. Such self-registration might be promoted in collaboration with selected publishers or information community partners as a means of the ISSN Network's strategic goal of increasing coverage of online resources beyond the limited staff capabilities of ISSN centers.

Other Potential Partnership Projects

The projects or organizations described below as potential partners for libraries in acquiring metadata which could form the basis of catalog records are all more or less in the formative stages as far as such collaborations are concerned. That many of these projects are still in the early stages presents both advantages and disadvantages. Although it is natural to want a well-established partner, it is also more difficult to bring about changes in the procedures of such partners. With partners whose operations are still under development, there will be more opportunities to influence development in ways mutually beneficial to all partners.

Before I discuss specific potential partners, I would like to mention a metadata framework--< indecs >--and a metadata standard--ONIX International--which are relevant to some of the potential partnership projects listed below. The < indecs > (interoperability of data in e-commerce systems) project "was created to address the need, in the digital environment, to put different creation identifiers and their supporting metadata into a framework where they could operate side by side, especially to support the management of intellectual property rights.... < indecs > is designed to help bridge the gap between the powerful but highly abstract technical models such as that expressed in the Resource Description Framework (RDF) and the more specific data models that are explicit or implicit in sector- or identifier-based metadata schemes." 9. The indecs metadata framework has influenced the DOI Foundation's metadata model as well as the development of ONIX and other metadata projects.

The ONIX International metadata set has been produced by EDItEUR jointly with the Association of American Publishers (Washington), Book Industry Communication (London), and the Book Industry Study Group (New York). ONIX Version 1.01 states that "ONIX International is the international standard for representing and communicating book industry product information in electronic form." 10. The ONIX metadata set includes product identifiers (ISBN, EAN-13, DOI, etc.); product descriptions (author, title, edition, publisher, publishing dates, series/set, subjects); product "promoters" (annotations, prizes, reviews); and product business data (supplier restrictions, pricing).

DOI (Digital Object Identifier)

According to the DOI Foundation's Web site, "The Digital Object Identifier (DOI) is an identification system for intellectual property in the digital environment. Developed by the International DOI Foundation on behalf of the publishing industry, its goals are to provide a framework to manage intellectual content, link customers with publishers, facilitate electronic commerce, and enable automated copyright management. " 11. Although early DOI registrations did not include associated metadata, according to the DOI Handbook (5.1), "Metadata is an essential component of the DOI System, and declaration of a limited "kernel" of metadata will soon become mandatory for all DOIs that are registered." 12. Additionally, "genres," defined areas of intellectual property, will be defined to serve different communities of interest and these genres will be able to define additional metadata elements appropriate to the particular genre.

Although the DOI has not as yet become a key identifier for Internet resources, the DOI syntax was approved by the American National Standards Organization in May 2000 and published as an official standard (Z39.8402000). According to the International DOI Foundation Annual Review dated September 2000, the DOI has gained in the number of participants and collaborations during the past year. 13. If the DOI continues to grow in use and participants, it might eventually prove a valuable source of metadata for Web-based resources. A potentially key application using the DOI is CrossRef, a collaborative effort which over 50 publishers had joined as of this writing (Oct. 2000). CrossRef enables linking between citations in one journal article to the cited content in another journal, even if that content is published by a different publisher and available on a different server. Success of CrossRef could give the DOI higher visibility as a viable system.

There is potential for collaboration between the DOI Foundation and libraries. The DOI Web site's FAQs indicate an interest in collaboration with libraries. The following question and answer opens the door to such collaboration:

> Question 2.4 Are there any plans to extend the DOI to bibliographic resources, like library card catalogues, to provide a subset of bibliographic information?
>
> The current plans for the prototype are for all participants in Internet-enabled publishing to determine how the DOI will work for their purposes, and we encourage other parties to begin considering the DOI as a tool for additional functions and services, such as metadata, bibliographic data and copyright management systems. 14.

Additionally, Norman Paskin, Director of the DOI Foundation, responded to this author's question about the Foundation's interest in potential collaboration with libraries for the purpose of sharing metadata with these words, "The concept you are presenting - that 'metadata created for other purposes such as copyright registration, ISSN registration, CIP applications, and identifier registrations could form a basis for... catalog records' -- is one which we strongly support, and which is in fact central to our efforts. Re-use of metadata is a natural consequence of current developments." 15. Paskin went on to cite several developments which would support such collaboration. First is the fact that the DOI Foundation is using principles from the indecs framework, and implementations like ONIX, as the basis of its metadata. Use of these standards would provide some common ground since LC's Network Development and MARC Standards Office is working with others on a "crosswalk" from ONIX to MARC21. Such a crosswalk would allow data supplied by publishers in connection with DOI registration to be converted into corresponding MARC21 fields. Additionally, the DOI is using the Handle System, which is the underlying system for the Copyright Office's CORDS project.

Currently the elements in DOI kernel metadata number seven: identifier, title, main creator and role, type, mode, genre. However, the DOI's support of ONIX's very rich metadata set might well result in the provision of many more elements by at least some publishers. Whether these elements would all be publicly available is a question, however, since the "kernel" metadata elements are the only elements which are definitely intended to be freely available as a look-up from the DOI. It will be up to each genre community to define access to other elements. However, even if such elements were not freely available to the public, libraries would be in a good bargaining position to negotiate access, perhaps in return for access to library authority files.

ISBN (International Standard Book Number)

Metadata collected in association with the registration of U.S. books for ISBN has long been published in Books in Print. Currently, ISBN agencies are in the process of exploring their future in the digital world, giving libraries the potential for another source of metadata for Web resources. In a position paper prepared by the U.S. ISBN Agency entitled, "The Digital World and the Ongoing Development of ISBN," 16. some guidelines for assigning ISBNs to digital files are listed. For example, "format/means of delivery are irrelevant in deciding whether a product needs an ISBN..." and "each format of a digital publication represents a new edition and should have a separate ISBN." There has been a recognition that some metadata elements applicable to print are not relevant for digital materials, and that digital materials may require the definition of new elements. Accordingly, the International ISBN Agency is in the process of determining core metadata elements appropriate to those materials produced using print on paper and those appropriate to digital products. This core metadata work is being done in conjunction with EDItEUR International

One problem for libraries is determining Web resources of potential value to their patrons. Since the ISBN is already in use by those publishers whose works libraries have traditionally collected, perhaps metadata collected in conjunction with ISBN registrations will provide one means of narrowing the ever-expanding universe of resources libraries have to consider for selection purposes. Another potential benefit of collaboration with ISBN is that the ISBN has both the advantages of a long-established system, as well as the advantages of being in the early stages of involvement with Web resources and thus potentially open to collaboration on metadata elements and their re-purposing as the basis for bibliographic records.

ISTC (International Standard Textual Work Code)

The International Standard Textual Work Code (ISTC) is Project 21047 under the auspices of the International Standards Organization (ISO), a step on the way to becoming an ISO standard. The purpose of the ISTC, as stated in its draft scope statement is, "to enable the efficient identification of and administration of rights to textual works, particularly in the digital information environment. The ISTC provides a means of uniquely identifying works of text in databases and other sources and for the exchange of information about those works among authors, agents, publishers, retailers, librarians, and other interested parties on an international level." 17.

It is important for our purposes to note that the ISTC is meant to be a work-level identifier, appropriate to all manifestations of the same work. Libraries have generally cataloged different manifestations of textual works on different bibliographic records. However, with the increasing proliferation of manifestations of works, many libraries are reconsidering this practice since patrons and reference staff find the multiplication of records for the same work confusing. This may be the beginning of a movement towards describing works --at least in some cases--rather than manifestations in library catalogs. Indeed, this is already becoming the case in some libraries which are following "one record policies" by simply adding URLs for online manifestations to existing records for print manifestation. So, metadata created in conjunction with ISTC registrations may be compatible with at least some library cataloging practices.

The ISTC will require registration by publishers or other interested parties. The form of such registrations has not yet been determined but it is not far-fetched to assume some form of Web-based registration might be offered. Metadata will be collected to support such registrations. The current project draft specifies the following metadata elements: title (at least one) with appropriate title type indicated; at least one author if on record, or if not, at least one contributor to the work with their respective roles indicated; whether or not the work is derived from another work and if so, the type of derivation; in the case of a derived work, the ISTC of the source work or the title if no ISTC exists for the source work; a unique identifier for the registration of the ISTC. The developers of the ISTC recognize that there will be a need to indicate the relationship between an ISBN or ISSN and the ISTC in various applications. That section of the draft is still under development.

The ISTC will be assigned through registration agencies. It is likely that multiple agencies may be established to serve various segments of the textual works community. Such registration agencies would be potential partners for bibliographic projects in their respective areas of focus.

OCLC Open Names Project

OCLC is developing a project called "Open Name Services." The premise of the project as stated on its preliminary Web site is that "Web services should be built around names and the communities that support them." On the Web site OCLC states that it is researching "how traditional names like ISBN can be used in more Web-based services and how these names can be used to link these services." 18. The initial focus is on ISBNs but the plan is to build similar services using a variety of names, such as--potentially- ISSN, SICI, Handles, etc.

The services associated with these names will have to be supported with metadata, much of which already exists. Because OCLC's base is the library community, the potential for collaborating in the collecting and sharing of metadata associated with the project would seem to be great. One of the project's supporting statements indicates "this would allow many of the traditional library services to be provided along with many new services." In fact, OCLC is actively soliciting the participation of the library community in this project, "We need groups like OCLC members and publishers to agree to open names. We then need organizations to step forward and commit to services on these names." Thus there exists potential for libraries to work together with OCLC on possible collaborations.

NSDP Web Template Study

In order to assess the usability of data supplied by publishers on registration templates, a study was carried out comparing unedited data by supplied publishers using NSDP's online ISSN application form with the completed CONSER serial records resulting from editing and updating by a cataloger in NSDP. The cataloger responsible for making assignments to electronic serials was asked to save "before" and "after" printouts for post-publication ISSN requests. At the time of the study, 220 records had been saved. A 25% random sample was taken, resulting in analysis of 55 records. Seven elements were chosen for comparison: Title, Variant Title, Frequency, Publisher, Place, Designation, and URL. A system was devised for scoring the data supplied by the publisher as either a "Match," "Close Match," or "No Match" when compared to data on the final cataloged record. One person did all of the scoring. "Match" constituted an exact match. "Close Match" was used for cases where there were only differences that would not affect searching or identification, such as capitalization, punctuation, or full form vs. abbreviated form. These differences were differences only in form-and minor ones at that-with no difference in fact. In the table below, "Match" and "Close Match" were added together to produce a combined score.

During the course of the scoring it became clear that the element, "Variant Title," presented scoring difficulties because sometimes publishers supplied variants that the cataloger did not include at all, while other times the cataloger did include the variant but constructed the variant in a different form. Because the scoring pattern for this element would not match that of the other elements, the element, Variant Title, was dropped.

The element with the highest percentage of exact matches was Frequency with 73% matches, and 16% close matches, for a combined score of 89%. In the case of Frequency, capitalization of the first letter of the frequency designation was ignored, and frequencies that differed only in capitalization were scored as matches. Although in some cases the difference between the data supplied by the publisher and that on the finished catalog record might be considered subtle by some, e.g., 6 times a year vs. bimonthly, this kind of difference was scored as "No Match" because in cataloging terms and in some serials check-in systems these are regarded as two different frequencies.

The element with the next highest number of matches was URL with 65% Matches and 24% Close Matches for a combined score of 89%. The Close Matches were mostly cases where the publisher supplied a tilde or spacing underscore and the cataloger had to convert those characters into their hex equivalents to be acceptable in the OCLC system. A new version of the NSDP conversion program now performs that conversion so today the percentage of exact matches would be 89%. The "No Match" cases occurred when the URL provided by the publisher included one or more typos, such as the use of capital I for the numeral 1; where the URL was not provided; or where the cataloger entered a URL specific to the serial while the publisher supplied a URL for the entire Web site on which the serial appeared.

The next highest combined score--82%--was for the Title element, surprisingly so, since serial catalogers have the perception that what the publisher considers to be the title often differs from what the cataloger considers to be the title. Although the combined score was relatively high, the "Match" score, 13%, was the lowest of any element because for this element only, capitalization was taken into account when determining "Matches." Capitalization of titles in catalog records does not follow standard grammatical rules but is nonetheless considered important for catalog record consistency and interpretation. Catalogers feel they must correct the capitalization supplied by publishers. Thus, for this element there were 69% "Close Matches." The 18% of the cases which fell into the "No Match" category consisted of cases where the publisher included what the cataloger considered to be a subtitle in the title field, or vice versa. The worst match, interestingly enough, was found on an application form for a serial with a generic title. The application form was completed by a monograph cataloger.

Place of publication resulted in 29% "Matches" and 51% "Close Matches" for a combined score of 80%. The Close Matches were usually the result of the publisher's inclusion of a full form of the place name which the cataloger abbreviated, or vice versa. Also, sometimes the place information supplied by the publisher varied in fullness from what resulted after editing by the cataloger. However, in 20% of the cases there was no match. Sometimes, no place was supplied on the form, in other cases the place supplied by the publisher was entirely different from the place the cataloger used, and in a few cases multiple places were given by the publisher while the cataloger chose only one.

Designation, the numbering or dating scheme used by the publisher to identify individual issues, had the next highest combined score: 75% , with 30% Matches and 45% Close Matches. Close Matches varied from what was supplied by the cataloger in the use of abbreviations and in whether enumeration or chronology or both were chosen. In one of the No Match cases, the publisher had supplied Vol. 1, no.1 as the designation of the first issue, whereas the cataloger edited the publisher's statement to read Vol. 1, no. 1-2. In other cases the caption was different, e.g., "issue" vs. "number." Cataloging rules require the designation to be transcribed as it appears on the publication.

Ironically the "publisher" element was the element with the lowest combined score: 44%, comprised of 40% Matches and 9% Close Matches. This high percentage of discrepancies resulted from different interpretations of the meaning of "publisher," especially for digital works. In several cases a personal publisher was given instead of a responsible corporate body. In other cases, multiple bodies were given when only one was chosen by the cataloger. In one case, a Web design company was given instead of the corporate body responsible for the content.

Ways to Obtain More Usable Data

Although information obtained from the publisher in the above study was often factually correct, and would have been acceptable in a typical database, the highly prescriptive cataloging rules in AACR2 as interpreted by the Library of Congress require even factually correct data to be edited by a cataloger for conformity to the rules. In many cases, the rules prescribe exact transcription of the data as it is presented on the publication. In cases where the transcription does not have to be exact, the ways in which what appears on the catalog record can differ from what appears on the publication (e.g., by being abbreviated, by omission of portions of the data) are also prescribed. Although these highly prescriptive rules seem to argue against the creation of even baseline catalog records by any automated means, the results of the above study would seem to indicate that records based on metadata supplied by publishers still show potential. There are various means by which data requiring less editing for general conformity to cataloging rules could be obtained. And, of course, there is also the potential for creating certain categories of records such as the "metadata records" which are the topic of this paper, which will be acknowledged as not following cataloging rules.

Following are some potential means for increasing the usability of publisher-supplied metadata:

1. Provide pull-down menus: for elements like Frequency where a list of MARC values exists, pull-down menus could be used to limit choices to those compatible with cataloging practice. Pull-down menus could also enable publishers to supply data which could be converted into MARC21 fixed field codes (country codes, codes to indicate current or ceased, start and end dates, etc.). NSDP catalogers now must add these codes to each record.
2. Include more specific instructions: for frequently misunderstood elements like "Publisher," clearer explanations can be given for how to supply the element. Examples taken from common situations could be given. Instructions could be provided for what to supply in cases of multiple publishers, or in the case of a personal and corporate publisher. Instructions could be given for how to treat Web designers.
3. Develop better conversion programs: just as the problem with tildes and underscores was solved by having the conversion program supply the correct hex values, programs could be designed to correct capitalization in titles or frequencies even if editing by the cataloger would still be required in some cases, particularly with capitalization in titles.
4. Provide interactive instructions: programs could be developed to "talk" a user through the application process, allowing him to ask for help, and giving feedback when certain responses were made.
5. Change certain cataloging rules: rules could be examined for practices out of sync with common practice, e.g., capitalization of words in titles, standardization of when to use abbreviations and which abbreviations to use regardless of how the word appeared on the publication.
6. Change descriptive cataloging practice: abandon description rooted in paragraphs on a catalog card where information not presented in certain places on resources is relegated to notes, in favor of a database approach where all information is of equal importance and values are supplied for a list of elements, akin to Dublin Core.

Preben Hansen of the Nordic Metadata Project is in the process of analyzing responses to a survey that has been available on the Web for almost three years soliciting input regarding use of the online Nordic Metadata template. Although specific results of the survey analysis are not yet available, Hansen indicated that the reactions to the template have been mostly positive. Comparison of Hansen's survey results with those obtained in the NSDP study is planned when those results become available.

Provision of Subject Data

The NISO draft standard Z39.85, Dublin Core Metadata Element Set, includes the following comment regarding the subject element, "Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme." 19. The unwieldy and even inappropriate results generated by most Internet search engines amply demonstrate the virtues of using controlled vocabularies for subject metadata.

However, subject analysis is one of the most expensive parts of the cataloging process and the one requiring the highest level of staff to perform it. In order to facilitate the provision of subject data using controlled vocabulary, the Nordic Metadata Project included in conjunction with its template, links to "all (as far as we know) vocabularies, e.g., controlled keyword lists, thesauri, classification systems, authority lists, general vocabulary system et cetera, which are freely available for navigation on the Web." 20. The template Web site includes a list of links to over 100 general and specialized tools, a list which would certainly frighten all but the most intrepid publishers. Even access only to the Library of Congress subject and name authority files were provided (these files became inaccessible-at least for the time being-when the Library's ILS went online in August 1999) publishers would still be faced with an enormous number of choices (5 large volumes worth) and a complex system of heading construction which it takes trained catalogers years to learn. For some time I have been advocating the development of a subset of LCSH which could be used as a tool by which publishers could supply subject metadata which would be fully compatible with LCSH. This would be a considerable undertaking but one with great potential benefit.

Another approach might be to use a subject list already in use by the publishing community. Such a list is the Book Industry Study Group's "BASIC Subject Heading Codes," a list which includes "over 2500 codes that may be assigned to books for bibliographic classification purposes, shelving in retail stores and searching online databases." 21. CIP's prototype New Books template includes access to the BASIC codes. Although the BASIC codes and the subjects they represent are more general than fully-subdivided LC subject headings, LCSH equivalents could most likely be determined for most of the BASIC codes. In this way, publishers could provide subject information using terms familiar to their industry while libraries could incorporate publisher-provided subject data into their catalogs using terms from their own authority files. An analysis of the potential of the BASIC codes to be converted into equivalent LCSH terms would be a useful study for a Library of Congress intern or research scholar to perform.

Finally, although this might seem to be testing the limits of what even the most motivated publishers might be inclined to provide, access to Library of Congress name authority files could also be made available to online template users-complete with basic instructions--so that, when possible, names of creators and contributors could be provided in authoritative form.

Conclusion: Call to Action

Libraries can no longer afford the luxury of acting as if they are the only organizations capable of describing resources. Catalog users, as studies have shown do not find significance in, or even understand, much of the information that catalogers labor to provide in a highly prescribed manner. The concept of exact transcription from a publication is particularly problematic for online resources which can change their appearance from one viewing to the next! Through the means described here and through other means suggested at this conference and elsewhere, libraries can and must make use of metadata created for other purposes to help bring some measure of bibliographic control to the ever-growing numbers of digital resources of interest to library patrons. Librarians need to share our expertise and help shape the development of metadata standards and metadata collection templates. Librarians need to share our name and subject heading expertise and authority files. Librarians need to collaborate to solve a common problem for the benefit of all in the publishing, library, and information communities. Librarians need to collaborate, not replicate. Librarians need to be partners, not competitors. There are more than enough resources to go around!

The potential partnerships described here are not intended to be exhaustive or prescriptive. Rather, they are meant to be illustrative of potential, some greater, some perhaps less. These examples are intended to provoke discussion and dialogue, which it is hoped will result in proposals that have the potential to result in control over a larger proportion of Web resources than might otherwise be possible using traditional means and resources.

What is offered here can be reduced to some take-away principles rather than a strict blueprint:

- A hierarchy of catalog record levels-with the lower levels based on publisher-supplied metadata--may be needed to bring Web-based resources under bibliographic control .
- Metadata created for other purposes can be re-purposed for library use.
- Resource creators and producers can create usable metadata, especially with librarians' help in developing standards, templates, guidelines, instructions-and motivation!
- Libraries cannot take sole responsibility for description of Web resources, but we can help, lead, guide and share expertise.

Every day thousands--if not tens of thousands--of new Web resources appear. Every day our invisible cataloging backlogs grow. The time to begin the "new order" is now!

NOTES

1. Committee on Information Technology Strategy for the Library of Congress, LC 21: A Digital Strategy for the Library of Congress. Prepublication copy. Washington, D.C.: National Academy Press, July 26, 2000, 5-6. http://www.nap.edu/books/0309071445.html
2. LC 21, 5-15.
3. Juha Hakala et al., Nordic Metadata Project. Final Report, Helsinki University Library, July 1998, 1. Introduction, http://www.lib.helsinki.fi/meta/nmfinal.htm
4. BIBLINK Project home page, http://hosted.ukoln.ac.uk/biblink/
5. Manjula Patel and Robina Clayphan, "Project BIBLINK: Linking Publishers and National Bibliographic Agencies." Proceedings: Internet Librarian 2000, March 2000 http://www.ukoln.ac.uk/metadata/publications/biblink/proj-biblink.html
6. LC 21, 3-7
7. Mary Levering, Associate Register of Copyrights, personal communication, Sept. 2000.
8. Levering, personal communication, Sept. 2000.
9. INDECS Framework, http://www.indecs.org
10. ONIX International, Version 1.01. http://www.editeur.org/onixfiles.html
11. DOI home page, http://www.doi.org
12. DOI handbook, 5.1. http://www.doi.org/handbook_2000/index.html
13. International DOI Foundation Annual Review, The Foundation, Sept. 2000. http://www.doi.org/publications.html
14. DOI home page, FAQ. a href="http://www.doi.org/publications.html
15. Norman Paskin, personal email October, 2000.
16. U.S. ISBN Agency, "The Digital World and the Ongoing Development of ISBN," New Providence, N.J., R.R. Bowker Company. http://www.bowker.com/standards/home/isbn/digitalworld.html
17. ISOTC 46/SC9 Working Group 3 for Project 21047: International Standard Textual Work Code, p. 1 http://www.nlc-bnc.ca/iso/tc46sc9/istc.htm
18. Open Name Services home page, http://names.oclc.org
19. NISO Draft Standard Z39.85-200X, Dublin Core Metadata Element Set (DCMES), http://www.niso.org/Z3985.html
20. Nordic Metadata Project. Final Report, p. 16.
21. BASIC Subject Heading Codes. http://www.bisg.org/subjectheadorder.html

---

Library of Congress

Comments: lcweb21@loc.gov (December 29, 2000)